Statistical Journal of the IAOS 32 (2016) 327–338
DOI 10.3233/SJI-161019
IOS Press

Evaluating a new proposal for detecting data falsification in surveys

The underlying causes of "high matches" between survey respondents

Katie Simmons*, Andrew Mercer, Steve Schwarzer and Courtney Kennedy
Pew Research Center, Washington, DC, USA

Abstract. A recent paper [1] proposed a new detection method for data falsification in surveys called the maximum percent match statistic. The statistic measures the maximum percentage of questions on which each respondent matches any other respondent in the dataset. The authors argue that valid survey data should have few respondents that match on more than 85% of questions. Based on this metric, the authors conclude that 1 in 5 publicly available international surveys contain data that is likely falsified. To evaluate this claim, we tested the sensitivity of the measure to variations in survey characteristics using: simulations on synthetic and survey data; evaluations of high quality domestic and international surveys with little risk of falsification; and regression analysis on 411 of Pew Research Center's international surveys. We find that the presence of high matches in a survey is extremely sensitive to natural, benign survey characteristics, such as the number of questions or number of response options. Our analysis indicates that the proposed metric is prone to generating false positives – suggesting falsification when, in fact, there is none. Thus, we find that the claim of widespread likely falsification based on this measure is not supported.

Keywords: Data quality, data falsification, duplicates, international surveys

1. Introduction

A pair of researchers recently proposed a new approach to detecting falsification in public opinion data [1]. The measure they introduce is the maximum percentage match statistic, which is the maximum percentage of questions on which each respondent matches any other respondent in the dataset. This metric is an extension of the traditional method of looking for respondents that are exact duplicates across all questions within a dataset. Using this measure, they argue that the presence of respondents that match another respondent on more than 85% of questions – what we refer to as a high match – indicates likely falsification. They apply this threshold to 1,008 international survey datasets and conclude that nearly one-in-five publicly available datasets contain data that is likely falsified [1, p. 9].

The claim that there is potentially widespread falsification in international surveys is clearly concerning. In this paper, we summarize an extensive evaluation of the proposed metric conducted by Pew Research Center, and find that the claim is not supported. Instead, we find that natural, benign survey features can explain high match rates. Specifically, the maximum percentage match statistic is extremely sensitive to the number of questions, number of response options, number of respondents, and homogeneity within the population. Under real-world conditions it is possible for respondents to match on any percentage of questions even when the survey data are valid and uncorrupted. Our analysis indicates the proposed threshold is prone to generating false positives – suggesting falsification when, in fact, there is none. Perhaps the most compelling evidence that casts doubt on the claim of widespread falsification is in the way the approach implicates some high-quality U.S. surveys. The thresh-

*Corresponding author: Katie Simmons, Pew Research Center, 1615 L Street NW, Suite 800, Washington DC 20002, USA. E-mail: ksimmons@pewresearch.org.

old generates false positives in data with no suspected falsification but that has similar characteristics to the international surveys called into question.

This paper proceeds as follows. First, we briefly review the problem of data falsification in surveys and how it is typically addressed. Second, we discuss our concerns about Kuriakose and Robbins' (K&R) proposed approach for identifying falsified data. Third, we outline the research steps we followed to evaluate the proposed metric and review in detail the results of our analysis. Finally, we conclude with a discussion of the findings and other ways the field is working to improve quality control methods.

2. Data falsification in surveys

Data falsification can occur at various levels of the survey process [2]. This includes, but is not limited to, field interviewers, supervisors, project managers and statisticians. Because falsification can take various forms, best practice is to have a variety of measures in place to mitigate the threat. The standard approach is twofold: prevention and detection [3–5]. Prevention includes developing a relationship with vendors; carefully training interviewers on the goals, protocols and design of a particular survey, as well as on the general principles and practices of interviewing; remunerating interviewers appropriately; limiting the number of interviews any given interviewer is responsible for; supervising a subset of the interviews for each interviewer; and, finally, re-contacting or re-interviewing, typically referred to as backchecking, a subset of the interviews of each interviewer to verify they were completed and conducted as documented. While these measures can be highly effective, they do not guarantee perfectly valid data [6,7].

Detection methods serve two purposes. First, they help to evaluate the performance of the costly prevention methods. Second, they can be used to identify falsified interviews that slipped past preventive measures [8–11]. Detection methods entail evaluation of key indicators, including paradata (interview length, timestamps, geocoding, timing of interviews), interviewer-related data (experience, daily workload, success rates), and interview-related data (characteristics of respondents, interview recordings, backchecking results), as well as analysis of the structure of responses (Benford's Law, refusals, extreme values, coherence of responses, consistency in time series, duplicates). But as with preventive measures, detection methods are not infallible. All concerns require intensive follow-up with vendors to determine the underlying explanation of the patterns found.

K&R propose a new detection method, suggesting a hard threshold for the number of high matches in a dataset to flag falsified data. The next section outlines our concerns about their proposal.

3. Overview of the high match measure

Given the challenges all researchers face in collecting high-quality survey data domestically and internationally, K&R's effort to develop a new diagnostic tool is part of an important line of research. However, the logic behind the authors' approach has two major flaws. The first is that the mathematical assumptions underpinning their argument are inappropriate. The second is that the analyses K&R use to validate their metric as an indication of falsified data are underspecified.

K&R's initial theoretical expectations about whether two respondents will give identical answers to a subset of questions (85%) are based on the likelihood of two respondents giving identical answers to all questions. The authors note that two respondents with a 95% chance of agreeing on each of 100 questions will match on all 100 questions less than 1% of the time [1, p. 4]. However, what the authors do not address is that the probability of matching on a subset of questions, such as 85%, is exponentially higher than the probability of matching on all questions. For example, in a 100-question survey, there is only one combination of answers such that two respondents match on all 100 questions. But there are 3.1×10^{17} different combinations of answers such that two respondents match on at least 85 of the questions. This means that two respondents with a 95% chance of agreeing on each of the 100 questions will agree on at least 85 of those questions more than 99% of the time.

In addition to using an 85% cutoff as an indicator for falsified data, K&R also suggest that the distribution of the maximum percent match statistic for a survey should resemble a Gumbel distribution. If the observed distribution deviates from the Gumbel, for instance due to additional modes or clumping toward the tail of the distribution, this is taken to be evidence of likely falsification. K&R do not provide any theoretical justification for expecting the maximum percent match statistic to follow a Gumbel distribution, and in fact there are several reasons why we should not expect this to be the case in general.

The belief that the statistic should follow any particular smooth distribution depends on the assumption that the values are independent and identically distributed. This assumption is wholly inappropriate for the maximum percent match statistic. For any respondent, the value of the statistic depends on how every other respondent answered each question in the survey, making its value entirely dependent on the other observations. More importantly, we should not expect the percentage of matching survey responses to be identically distributed. Within any national population, we should expect to observe not a single distribution, but a mixture of distributions corresponding to the different subpopulations. Distinct subgroups within a surveyed population will share different numbers of characteristics or opinions that are measured in the survey and that essentially represent different data generating processes. Particularly homogenous subgroups within the population will appear as additional modes or bunching toward the tail of the distribution. As with the 85% cutoff, the probability of observing deviations from the expected distribution proposed by K&R will depend greatly on the characteristics of the specific survey and the target population.

These critiques point to the larger weakness in the approach taken by K&R – namely that the authors do not systematically evaluate the survey characteristics that would cause the number of high matches to vary, such as the sample size, the number of questions, the number of response options or homogeneity within the population. These parameters have a direct bearing on both the number of possible response combinations and the number of respondents that potentially match.

K&R assert that their Monte Carlo simulations provide a conservative estimate of the distribution of the maximum percent match statistic. As we will show, however, they chose a particular set of conditions for their simulations – 100 questions, 1,000 respondents, 0.5 means for all variables – that led them to find few high matches. In particular, the assumption that all variables have a mean of 0.5 bears little resemblance to reality. In most public opinion surveys, some proportions are closer to either zero or one, reflecting the fact that there are often majority opinions or behaviors on topics studied in surveys. Assuming that the mean of each and every question in a survey is 0.5 underestimates the degree to which there is some natural similarity between respondents.

To fully understand whether the presence of high matches in a survey dataset is a result of fraud or of various survey characteristics, we pursued a multistep research design described in the following section.

Table 1
Expected effect of parameters on percentage of high matches

	As parameter:	% of high matches should:
Number of questions	Increases	Decrease
Number of response options	Increases	Decrease
Number of respondents	Increases	Increase
Homogeneity in population	Increases	Increase

4. Evaluating the measure

We evaluated the sensitivity of the proposed measure to additional parameters not tested in the original paper in an attempt to better understand how the statistic would react to variation in real-world survey conditions. The first parameter we analyzed is the number of questions. With more questions, the probability that two respondents match on a large percentage of those questions should decline. The second is the number of response options in the questions. With more response options, respondents are less likely to give the same answer as someone else. The third is the number of respondents. With more respondents in the dataset, there are more opportunities for respondents to match. The fourth is the homogeneity within the sample. When the content of the survey or the population being surveyed lead to greater homogeneity of opinion, either in the full sample or among certain subgroups, the probability of a match between two respondents should increase. Table 1 summarizes these expectations.

We evaluated the impact of these four parameters on the percentage of high matches in datasets with simulations using synthetic data and survey data, as well as with analysis of high-quality U.S. and international surveys. We find that K&R's measure is extremely sensitive to all four parameters discussed above.

4.1. Simulations with synthetic data

Simulations are useful because they allow the researcher to conduct analysis in a very controlled environment. We can set the conditions for the parameters we think should matter and evaluate how a statistic changes when we vary just one of those parameters. This type of analysis allows us to develop theoretical expectations about how real-world data should behave. A serious limitation of using synthetic data for this type of analysis, however, is that if the assumptions differ substantially from real-world situations, the theoretical expectations derived from them may not be very useful.

K&R simulated surveys consisting of 1,000 respondents and 100 independent binary variables, each with

a mean of 0.5. We extended their analysis by varying each of the following parameters. For the number of questions, we tested values ranging from 20 to 120 in increments of 20. For the number of respondents, we tested values from 500 to 2,500 in increments of 500. We conducted this set of simulations twice. The first time we set the mean of each variable at 0.5, consistent with K&R's approach. The second time, we set the mean of each variable at random from a uniform distribution between 0 and 1. This second condition more closely resembles the reality of survey data, where some variables have means close to 0.5 while others have means that approach the extremes of either 0 or 1. Variables with means closer to 0 or 1 represent the type of questions on surveys where respondents are more homogeneous in their opinions.

For each simulated survey, we calculated the proportion of respondents classified as a high match, meaning the respondent matches another respondent on more than 85% of questions. We replicated each combination of sample size and number of respondents 1,000 times.

We find that when the variable means are fixed at 0.5, there are no respondents classified as a high match in any of the simulations with 100 or more questions, and only a handful meet the 85% threshold with 40 or 60 questions, regardless of the sample size. Only at 20 questions do a substantial percentage of respondents qualify as high matches, with a median of 10% when the sample size is 500 and a median of 40% when the sample size is 2,500. The results for the datasets with 100 variables and 1,000 respondents are consistent with K&R's simulation (The graph for all of these simulations is in Fig. 4).

However, when the variable means are allowed to vary randomly, a very different picture emerges. Figure 1 compares the results of the two sets of simulations when the sample size is fixed at 1,000 (see Fig. 5 for the results of all simulations). When the means vary across questions, the proportion of respondents that qualify as high matches increases dramatically. With 20 questions, the median survey has 91% high matches, while at 60 questions, the median survey has 15%. Even at 120 questions, more than one-third of the simulations have high matches, ranging from 2% to 14%.

In their simulations, K&R tested a single combination of survey parameters – 1,000 respondents and 100 binary questions with means implicitly fixed at 0.5. Our additional simulations demonstrate that their results are highly sensitive to their choice of parameters.

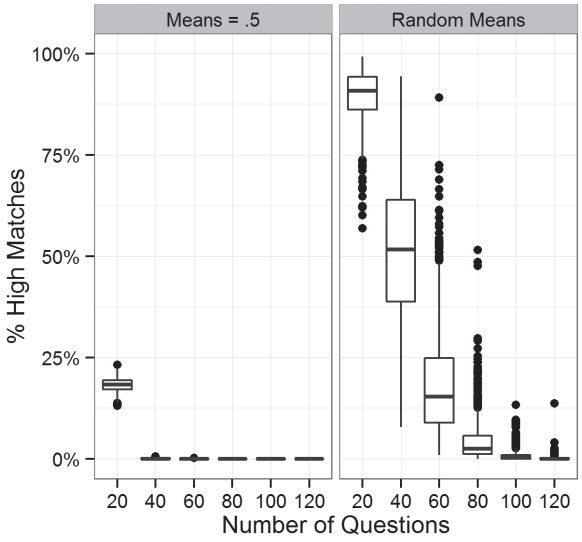

Fig. 1. Sensitivity of high matches to number of questions and means. Box plots of distribution of the percentage of respondents with more than 85% matching responses over 1,000 simulations for $n = 1{,}000$. Simulated datasets consist of independent, randomly generated, binary variables with means of 0.5 and means randomly assigned from a uniform distribution. Each combination of sample size and number of questions was simulated 1,000 times.

Surveys with fewer questions, larger samples or items with high levels of respondent agreement can all be expected to produce respondents who are more similar to one another. Furthermore, these synthetic data simulations remain highly unrealistic: Questions only have two response categories and they are all independent. This is not an adequate basis for generating hypotheses about what should be expected in practice, as questions are often correlated with one another and frequently include more response options.

4.2. Simulations with survey data

In order to replicate more realistic survey conditions while still retaining control over the features of the survey, we also conducted simulations with survey data to understand the impact of various parameters in real-world conditions. We used the 2012 American National Election Study and the Arab Barometer Wave III Lebanon surveys as the basis for additional simulations. These are two high-quality surveys that, based on K&R's approach, are assumed to be free of duplication. The two surveys have large sample sizes, with 2,054 and 1,200 cases respectively, and lengthy questionnaires, with at least 200 substantive

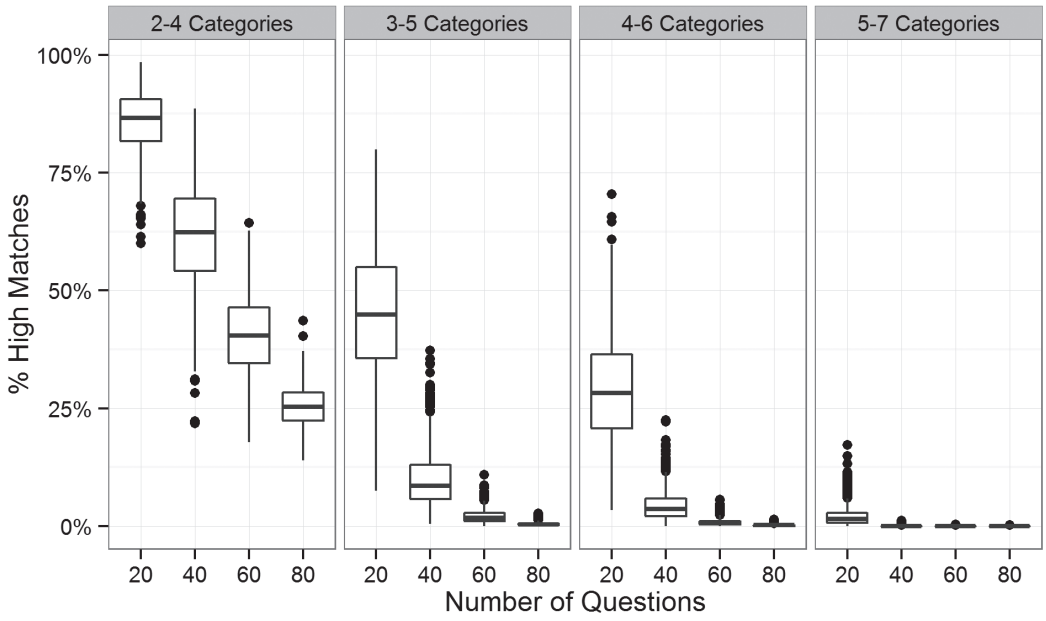

Fig. 2. Sensitivity of high matches to number of response categories. Box plots of distribution of the percentage of respondents with more than 85% matching responses over 1,000 simulations for $n = 1{,}000$. Simulated datasets are drawn from the 2012 American National Election Studies pre-election survey. Each combination of sample size, number of questions and number of response categories was simulated 1,000 times.

questions.[1] The size of the surveys allowed us to randomly select subsamples of questions and respondents from all questions and all respondents available. By doing so, we were able to vary key parameters in a semi-controlled environment using real-world survey data where the variables and respondents are correlated. We excluded any questions for which more than 10% of respondents have missing values.

We also evaluated the impact of the number of response options in the questions using the ANES. We did this by performing simulations that varied the number of response options per question in addition to the number of questions and the sample size. Rather than randomly select from all possible questions in a survey, these simulations randomly selected from questions that have two to four, three to five, four to six or five to seven response categories.

First, we present the results from the ANES data to assess how the share of high matches in a survey is related to the number of response categories in survey questions. Figure 2 contains the results for the datasets with 1,000 respondents.

As with the synthetic simulations, the number of questions and respondents continue to have an impact on the percentage of high matches. We also find that as the number of response options decreases, the percentage of high matches increases considerably. As expected, this also varies with the number of questions and the sample size, but when there are only two to four response options, the median percentage of high matches ranges from 87% when there are 20 questions to 25% when there are 80 questions. This confirms what we would expect intuitively – that the proportion of high matches in a survey will be sensitive not only to the number of questions, but also to the types of questions included in a survey. Most surveys will include a mix of questions with different numbers of response options ranging from few to many. For any given survey, the details of that distribution are another important determinant of the number of high matches that would be present.

The results for the two to four response options also represent a significant departure from the results obtained with the synthetic data simulations that replicated K&R's approach. With the synthetic data, when the number of respondents is 1,000, the variable means are fixed at 0.5, the number of questions is 80 and the number of response options is two, there are no high matches under these conditions. Under the same conditions in the ANES (with the exception of 0.5 means),

[1] For all datasets, we only analyzed substantive variables – meaning no demographics and no paradata – and we only included variables for which less than 10% of the sample was not asked the question. This approach was to be consistent with K&R's analysis.

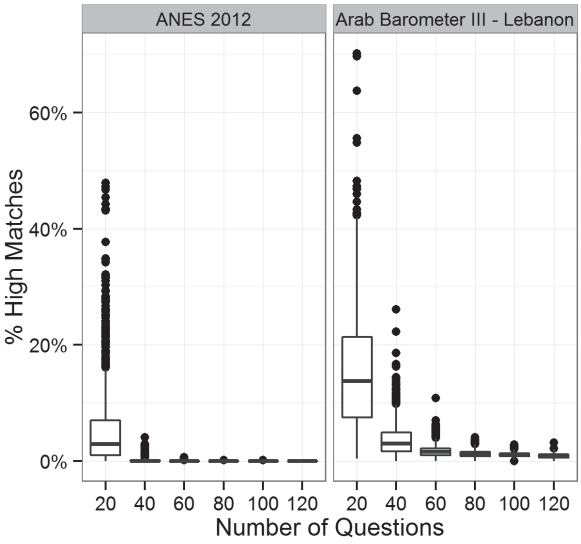

Fig. 3. Comparison of high matches in the ANES and Arab Barometer simulations. Box plots of distribution of the percentage of respondents with more than 85% matching responses over 1,000 simulations for $n = 1{,}000$. Simulated datasets are drawn from the 2012 American National Election Studies pre-election survey and the Arab Barometer III Lebanon survey. Each combination of sample size and number of questions was simulated 1,000 times.

the median percentage of high matches across 1,000 replications is 25%. This comparison re-emphasizes that, contrary to K&R's assertion, their simulations are not a conservative estimate of the percentage of high matches in real-world survey data. Instead, this comparison suggests that a threshold based on simulations with synthetic data is not relevant for what we should see in real-world data.

We also conducted comparable simulations with the Arab Barometer Wave III Lebanon survey, which was fielded in 2013. The purpose of this comparison was to evaluate the presence of high matches under various conditions in a non-falsified dataset that surveyed a non-U.S. population. Figure 3 contains a comparison of simulations drawn from the ANES and the Arab Barometer surveys with a sample size of 1,000 and varying the number of questions between 20 and 120 questions.[2] In this set of simulations, the number of response options in the questions is allowed to vary.

We see very different distributions of high matches in the ANES and Arab Barometer surveys. Whereas the percentage of high matches in the ANES is nearly zero for all but the 20-question condition, the Lebanon simulations reflect a higher proportion of high matches, even at 100 or 120 questions. This indicates that the probability of any two respondents matching on more than 85% of questions depends not just on the number of respondents or the number of questions, but also on the particular survey content and the population being surveyed. In other words, a threshold based on the ANES and other surveys conducted in the United States does not necessarily generalize to other countries. Even within a single country, there is no a priori reason to believe the distribution of high matches observed on one survey should be similar to another survey with different content.

[2]We tested additional larger and smaller sample sizes and larger numbers of variables; the results are consistent with those shown here.

4.3. Pew Research Center U.S. survey data

The level of homogeneity among respondents in any given survey depends on both the content of the survey and the population being surveyed. To evaluate the impact of these two components, we first analyzed the level of homogeneity across different surveys of the same population, in this case the U.S. general adult population. We then analyzed the level of homogeneity across different populations – various U.S. partisan and religious subgroups – on the same set of survey questions.

For the evaluation of the content of the survey, we compared the percentage of high matches in domestic survey data from Pew Research Center with the theoretical expectations derived from the simulations based on the ANES. Like the ANES, the survey data analyzed here present little concern about the presence of falsified data. The data come from random-digit-dial telephone surveys with centralized and live interviewer monitoring and the collection of detailed contact data. Unlike the ANES, the surveys tend to have shorter questionnaires that are focused on a few specific topics. For the analysis, we reviewed four political surveys conducted by Pew Research Center in 2014 and 2015, including the large 2014 Political Polarization and Typology survey, an October 2014 election survey and two typical monthly surveys from 2015 that covered major political issues in the news at the time.

For the evaluation of different populations, we used the four political surveys described above as well as the 2014 Religious Landscape Study, which is a nationally representative telephone survey of 35,071 U.S. adults with 41 substantive questions asked of all respondents. Data collection for the Landscape Study was conducted by three different research firms. In general, the American population is very diverse. But it also

includes more homogeneous subgroups with respect to different issues covered by each survey. The political surveys ask about a range of issues that polarize Democrats and Republicans, enabling us to evaluate how the percentage of high matches differs among partisan groups. The Religious Landscape Study includes, among other things, questions on religious identity and beliefs and practices. The large size of the survey allowed us to analyze religious groups that are relatively small, homogeneous segments of the population, such as Mormons, with a robust sample size.

4.3.1. Evaluating the impact of questionnaire content

The four political surveys we analyzed have a relatively modest number of questions asked of the entire sample, ranging from 29 to 52. The number of respondents ranges from 1,502 to 2,003 for the three monthly surveys and is 10,013 for the Polarization study. Table 2 reports the percentage of respondents that match another respondent on more than 85% of substantive variables for each of the four surveys analyzed, along with the parameters for each survey, including number of respondents, number of questions and percentage of questions with five or more response options.

Overall, across the four surveys, there are substantial percentages of high matches, ranging from 12% in the September 2015 survey to 39% in the 2014 Polarization study. In large part, the number of high matches is driven by the low number of questions typically asked, the relatively low number of response options and the large sample sizes, especially in the Polarization study.

Nonetheless, in the July 2015 survey with 52 questions and 2,002 respondents, we find 13% of the sample is a high match. In the simulations with synthetic data with 0.5 means, as well as the simulations with the ANES data, the median percentage of high matches across 1,000 replications with these conditions is 0. Given that there is little concern about the presence of data falsification in the July 2015 survey, this comparison reveals that the content and context of the questionnaire can have a significant impact on the percentage of high matches in a dataset. The findings also suggest that a single threshold for the maximum percent match statistic based on simulations with synthetic data and the ANES may not be appropriate.

4.3.2. Evaluating the impact of population homogeneity

To understand the effect of population homogeneity on the percentage of high matches in a dataset due to subgroup agreement, we evaluated how the percentage of high matches varies by partisan group in the four political surveys. Table 3 shows the percentage of respondents in each partisan group for each survey that is a high match.

People who identify with a political party tend to be more polarized and firm in their political beliefs than those who say they are independent, and therefore we expect higher levels of homogeneity among partisans. Indeed, we find that Republicans and Democrats tend to have higher percentages of high matches than independents, though the exact percentage varies by survey.

We also investigated the impact of population homogeneity using the 2014 Religious Landscape Study. Since the percentmatch tool developed by K&R is unable to process a dataset of this size, we evaluated 10 random samples from the dataset of roughly 1,000 respondents each to get a sense for the number of high matches overall. The highest percentage of respondents that matched another respondent on more than 85% of the substantive variables in any of the 10 random sub-samples was 6%. In addition, we analyzed random samples of approximately 1,000 respondents for each of the three fieldhouses that conducted the survey. The fieldhouses exhibited similar percentages of high matches, ranging between 4% and 7%, bolstering the argument that these data are not falsified.

Once we look at specific religious subgroups, however, the percentage of high matches increases considerably. We analyzed four religious subgroups separately using the same set of 41 questions. In this set of 41 questions, 54% of the questions have five or more response options. Table 4 lists the percentage of high matches and number of respondents for each of the four different religious groups. Mormons have the highest percentage, with 39% of respondents that are a high match. Atheists have 33% high matches and Southern Baptists have 31% high matches. On many religion surveys, these three religious groups tend to be more homogeneous in their beliefs and practices than other American religious groups. Jews, on the other hand, have very few high matches (1%). As with the partisan differences on the political survey, the religious differences on this survey suggest that homogeneity within specific populations can drive up the percentage of high matches in the dataset without indicating the presence of falsified data.

The findings from both the political surveys and the Religious Landscape Study indicate that even in high-quality datasets in the U.S. conducted under rigorous quality controls, there is considerable variation in the

Table 2
High matches in U.S. political surveys

	% High matches	Sample size	Number of questions	% of questions with 5+ resp. options
September 2015	12	1,502	32	50
July 2015	13	2,002	52	37
October 2014	24	2,003	29	48
Polarization 2014	39	10,013	36	14

Pew Research Center surveys, conducted between January 2014 and September 2015.

Table 3
High matches by partisan group

	Republican	Independent	Democrat
September 2015	21	7	10
July 2015	8	7	24
October 2014	39	14	25
Polarization 2014	42	36	43

Percentage of high matches by party self-identification. Pew Research Center surveys, conducted between January 2014 and September 2015.

Table 4
High matches among religious groups in Landscape Study

	% High matches	Sample size
Mormons	39	645
Atheists	33	1,098
S. Baptists	31	1,845
Jews	1	850

Religious Landscape Study, 2014.

percentage of high matches. This variation is driven in part by the topics covered by the survey and the homogeneity of the population with respect to those topics. The ANES surveys are conducted with a very diverse population using a varied and long questionnaire. The findings in this section, along with the results of the simulations discussed earlier, suggest that the distribution of the maximum percentage match statistic in the ANES is not generalizable to other surveys or other populations.

4.4. Pew Research Center international data, 2002–2013

Finally, we tested the relationship between survey characteristics and the percentage of high matches using 411 publicly available datasets that were part of a cross-national project between 2002 and 2013 from Pew Research Center. The data were drawn from international surveys conducted in all regions of the world. The Center's Global Attitudes surveys, which have been conducted yearly in roughly 20 to 40 countries since 2002, account for 337 surveys in this analysis. The content of the questionnaire varies somewhat from year to year and country to country, but a significant portion of the questions relate to foreign affairs and attitudes about the United States. The Center's international research on religion accounts for the other 74 surveys in this analysis. These questionnaires focus primarily on religious beliefs and practices, as well as attitudes about morality and the role of religion in society.

For each dataset, we have four key variables: the total number of substantive questions asked on the survey, the percentage of questions with five or more response options, the total sample size, and whether the survey focused on religious beliefs and practices. The first three variables are relatively straightforward measures of the parameters discussed above. We expect the number of questions and the percentage of questions with five or more response options to have a negative effect on the percentage of high matches on a survey, and the total sample size to have a positive effect. The religion survey variable is a blunt measure to try to capture the effect of questionnaire content on homogeneity. The religion surveys conducted internationally have asked questions about basic religious beliefs and practices, such as whether a respondent believes in God or how often someone prays, in many countries that are relatively homogeneous in their religious composition and/or are relatively devout. Given this, we expect the religion surveys to exhibit a higher level of homogeneity than the Global Attitudes surveys, which tend to ask questions that exhibit less agreement among the population in many countries.

To test this, we ran a regression analysis of the relationship between the four parameters and the percentage of high matches across all 411 surveys. In the simulations, the relationship between the number of questions and the percentage of high matches is clearly nonlinear. Because of this, we included a quadratic term for the number of questions in the regression model.

Table 5 shows the results of the analysis.[3] Consistent with our expectations, the percentage of high matches

[3]The range of the variables is as follows. The percentage of high matches ranges from 0 to 92%, with an average of 8%. The number of questions ranges from 5 to 157, with an average of 87. The per-

Table 5
Influence of survey characteristics on the percentage of high matches in international surveys

Variable	Estimate	Std. Err	p-value
Number of questions	−0.781	0.072	< 0.001
Number of questions $^{\wedge 2}$	0.004	0.0004	< 0.001
Percent of questions with 5+ response options	−0.485	0.053	< 0.001
Number of respondents	0.003	0.001	0.012
Religion survey indicator	3.851	1.747	0.028
Intercept	68.412	4.004	< 0.001

Dependent variable is the percentage of respondents who match on more than 85% of variables. Data are drawn from 411 Pew Research Center international surveys. Adjusted R^2 for the model is 0.401.

decreases exponentially as the number of questions increases. We also find that surveys with a higher percentage of questions with five or more response options have significantly lower percentages of high matches. Meanwhile, the number of respondents in the survey significantly increases the presence of high matches. Finally, we find that religion surveys are more likely to have a higher percentage of high matches than the surveys that primarily cover foreign affairs and other topics.

Taken together, these results provide further evidence that the presence of high matches in a survey dataset is heavily influenced by benign characteristics of the survey.

5. Discussion

K&R assert in their paper that two respondents that match on a high percentage of questions should be a rare occurrence in valid data, and that the presence of respondents that match on more than 85% of questions is an indication of likely falsification. They make their case for this conclusion based on a review of public opinion literature, simulations with synthetic data and analysis of data from the American National Elections Study and the General Social Survey.

However, the assumptions underpinning their argument – and the datasets they used to develop their threshold – raise some serious questions about whether high matches in a dataset are an indicator of likely falsification or whether high matches may result from various permutations of the characteristics of the survey. The goal of this paper was to understand the conditions under which high matches may be present in valid survey data.

Using synthetic simulations as well as high-quality domestic and international datasets, we show that the percentage of high matches varies widely across datasets and is influenced by a variety of factors. The characteristics of a survey, such as the number of questions, the number of response options, the number of respondents and the homogeneity of the population, or subgroups therein, all affect the percentage of high matches in a dataset. The results show that it is possible to obtain any value of the maximum percent match statistic in non-falsified data, depending on the survey parameters. Thus, setting a threshold for the statistic and applying it uniformly across surveys is a flawed approach for detecting falsification. In fact, eliminating respondents from a dataset based on this measure may introduce selection bias into survey data and serve to reduce data quality, rather than improve it.

The sensitivity of K&R's measure to these characteristics highlights the need to understand the study-specific environment of a survey to evaluate the meaning of any statistical assessment of the data. Bredl et al. [13] note that "one has to keep in mind that striking indicator values are not necessarily caused by data fabrication but may also be the result of "conventional" interviewer effects or cluster-related design effects [spatial homogeneity]" (p. 20). Judge and Schechter [14] conclude from their analysis of survey data that multiple factors might contribute to suspicious-looking patterns in data and that detection methods should not be used "in isolation when judging the quality of a dataset" (p. 24). Any data quality assessment needs to take into account the specific design characteristics and the specific conditions of a survey before drawing conclusions.

Nonetheless, K&R are taking part in an important discussion about how to improve detection methods for data falsification. The use of new technologies for face-to-face surveys, such as devices for computer-assisted personal interviewing, present many new possibilities when it comes to ensuring data quality through prevention and detection methods.

centage of questions with five or more response options ranges from 20% to 100%, with an average of 67%. The number of respondents ranges from 485 to 4,018, with an average of 1,069.

One especially promising innovation is the measurement of time throughout the survey in face-to-face studies. This includes the overall length of a survey, from start to finish, but also the time it takes to go through sections of the questionnaire, or to answer a specific question. The measurement of section timings can be used to evaluate whether the respondent or interviewer may have had unusual difficulties with a particular section, or whether the interviewer may not have taken the appropriate amount of time to ask certain questions. Another promising avenue for detection of falsified data is the use of computer audio-recorded interviewing to record random points in the interview. This allows the researcher to review whether the respondent and/or interviewer were speaking and whether the same respondent was answering the questions throughout the survey. Other aspects that could be efficiently embedded in a computer-assisted interviewing environment are within household selection procedures, as well as the collection of geographical tracking information. The community is still exploring how to use these new tools in the most effective way. Regardless, technological advances make it much easier to collect data on important aspects of the survey process beyond substantive data (i.e. paradata or auxiliary data). These data can be converted from a byproduct of the survey into a primary analytical tool for assessing survey quality.

Still, even these new approaches would need to be evaluated along with a variety of other indicators. Any statistical analysis of data quality has its limitations. Thus, researchers should try to involve vendors in the assessment of the quality of the data.

Engaging vendors provides two benefits. First, it helps to reduce the information gap created by the principal-agent dilemma by allowing researchers to learn something about the specific conditions under which interviewers were operating. This will contribute to the overall interpretation of the data itself, but will also help with the evaluation of suspicious data patterns. Second, involving vendors closes the circle of prevention and detection and places the whole assessment in the wider context of quality assurance. The involvement of vendors allows the vendor and the researcher to evaluate and learn for future projects. The findings from detection measures should inform the design and structure of future questionnaires, lead to new approaches to incentivize interviewers, and assist with the development of new prevention and detection methods.

Acknowledgements

We wish to thank the three outside researchers who provided excellent feedback on our paper prior to submission. We also are grateful to the following for their contribution to and comments on our paper: Becka Alper, James Bell, Jill Carle, Danielle Cuddington, Claudia Deane, Michael Dimock, Meredith Dost, Elizabeth Gross, Peter Henne, Ruth Igielnik, Scott Keeter, David Kent, Dorothy Manevich, Besheer Mohammed, Baxter Oliphant, Bridget Parker, Dennis Quinn, Aleksandra Sandstrom, Anne Shi and Richard Wike.

References

[1] N. Kuriakose and M. Robbins, Falsification in survey research: Detecting near duplicate observations, *Statistical Journal of the IAOS* Forthcoming 2016.

[2] R. Groves, F. Fowler, M. Couper, J. Lepkowski, E. Singer and R. Tourangeau, Survey methodology. Hoboken, N.J.: Wiley; 2009.

[3] Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects [Internet]. AAPOR; 2003 Apr 21. Available from: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf.

[4] L. Lyberg and P. Biemer, Quality Assurance and Quality Control in Surveys, in: *International handbook of survey methodology*, E. de Leeuw, J. Hox and D. Dillman, eds, New York: Lawrence Erlbaum Associates, 2008.

[5] L. Lyberg and D.M. Stukel, Quality Assurance and Quality Control in Cross-National Comparative Studies, in: *Survey methods in multinational, multiregional, and multicultural contexts*, J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler et al., editors. Hoboken, NJ: Wiley; 2010.

[6] A. Koch, Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. ZUMA Nachrichten [Internet]. 1995. Available from: http://www.ssoar.info/ssoar/handle/document/20898.

[7] C. Hood and J. Bushery, Getting more bang from the reinterview buck: Identifying 'at risk' interviewers [Internet]. Proceedings of the American Statistical Association; 1997. Available from: http://www.amstat.org/sections/srms/Proceedings/.

[8] S. Bredl, P. Winker and K. Kötschau, A statistical approach to detect interviewer falsification of survey data, *Survey Methodology [Internet]* **38**(1) (June 2012), 1–10. Available from: http://www.statcan.gc.ca/pub/12-001-x/2012001/article/11680-eng.pdf.

[9] S. Diakité, Statistical methods for the detection of falsified data by interviewers and application survey data in Africa [Internet], Sixth International Conference on Agricultural Statistics; 2013. Available from: http://www.statistics.gov.hk/wsc/CPS005-P7-S.pdf.

[10] N. Menold and C. Kemper, How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys, *International Journal of Public Opinion Research [Internet]* **26**(1) (2014), 41–65. Available from: http://ijpor.oxfordjournals.org/content/26/1/41.abstract?sid=e1af9146-4073-418d-914f-e61f4971fc1b.

[11] P. Winker, N. Menold, N. Storfinger, C. Kemper and S. Stutkowski, A Method for ex-post Identification of Falsifications in Survey Data [Internet]. NTTS 2013 – Conferences on New Techniques and Technologies for Statistics; 2013. Available from: http://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_93.pdf_en.

[12] F. Benford, The law of anomalous numbers, *Proceedings of the American Philosophical Society [Internet]* **78**(4) (31 Mar 1938), 551–572. Available from: http://www.jstor.org/stable/984802.

[13] S. Bredl, N. Storfinger and N. Menold, A literature review of methods to detect fabricated survey data. Discussion Paper from Justus Liebig University Giessen, Center for International Development and Environmental Research (ZEU) [Internet]. 2011. Available from: https://www.econstor.eu/bitstream/10419/74449/1/746858302.pdf.

[14] G. Judge and L. Schechter, Detecting problems in survey data using Benford's law, *The Journal of Human Resources [Internet]* **44**(1) (2009), 1–24. Available from: http://jhr.uwpress.org/content/44/1/1.refs.

[15] The American National Election Studies (ANES), The ANES 2012 Time Series Study [data file]. Stanford University and the University of Michigan: Stanford, CA and Ann Arbor, MI; 2012. Available from: http://www.electionstudies.org/.

[16] M. Tessler, A. Jamal, M. Shteiwi, K. Shikaki, M. Robbins, R. Hamami et al., Arab Barometer: Public Opinion Survey Conducted in Lebanon, 2012–2014 [data files]. Ann Arbor, MI; 2015 Nov 31. Available from: http://doi.org/10.3886/ICPSR36273.v1.

[17] 2014 Political Polarization Survey [data file]. Pew Research Center: Washington, DC. Available from: http://www.people-press.org/category/datasets/page/2/?download=20057011.

[18] October 2014 Political Survey [data file]. Pew Research Center: Washington, DC. Available from: http://www.people-press.org/category/datasets/?download=20056960.

[19] July 2015 Political Survey [data file]. Pew Research Center: Washington, DC. Available from: http://www.people-press.org/category/datasets/?download=20059299.

[20] September 2015 Political Survey [data file]. Pew Research Center: Washington, DC. Available from: http://www.people-press.org/category/datasets/?download=20059305.

[21] 2014 Religious Landscape Study [data file]. Pew Research Center: Washington, DC.

[22] Global Attitudes Survey datasets 2002–2013 [data files]. Pew Research Center: Washington, DC. Available from: http://www.pewglobal.org/category/datasets/.

[23] Spirit and Power: A 10-Country Survey of Pentecostals, 2006 [data files]. Pew Research Center: Washington, DC. Available from: http://www.pewforum.org/datasets/spirit-and-power-a-10-country-survey-of-pentecostals/.

[24] The World's Muslims, 2008–2012 [data files]. Pew Research Center: Washington, DC. Available from: http://www.pewforum.org/datasets/the-worlds-muslims/.

[25] Religion in Latin America, 2013–2014 [data files]. Pew Research Center: Washington, DC. Available from: http://www.pewforum.org/datasets/religion-in-latin-america/.

Appendix

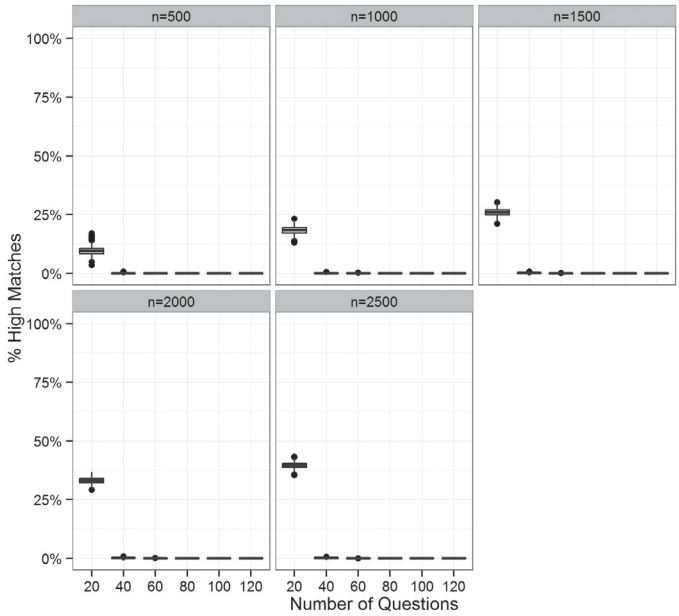

Fig. 4. Synthetic data simulations with mean fixed at 0.5. Box plots of distribution of the percentage of respondents with more than 85% matching responses over 1,000 simulations. Simulated datasets consist of independent, randomly generated, binary variables with means of 0.5. Each combination of sample size and number of questions was simulated 1,000 times.

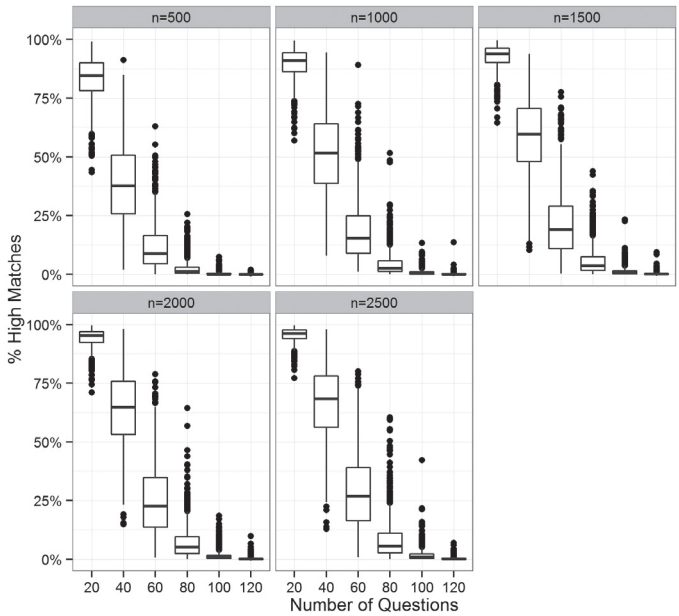

Fig. 5. Synthetic data simulations with variable means. Box plots of distribution of the percentage of respondents with more than 85% matching responses over 1,000 simulations. Simulated datasets consist of independent, randomly generated, binary variables with randomly assigned means of between 0 and 1. Each combination of sample size and number of questions was simulated 1,000 times.