Journal of Survey Statistics and Methodology (2020) ***00****, 1–25*

RELATIONS BETWEEN VARIABLES AND TRENDS OVER TIME IN RDD TELEPHONE AND NONPROBABILITY SAMPLE INTERNET SURVEYS

JOSH PASEK*
JON A. KROSNICK

Survey researchers today can choose between relatively higher-cost probability sample telephone surveys and lower-cost surveys of nonprobability samples of potential respondents who complete questionnaires via the internet. Previous studies generally indicated that the former yield more accurate distributions of variables, but little work to date has explored the impact of mode and sampling on associations between variables and trends over time. The current study did so using parallel surveys conducted in 2010 focused on opinions, events, behavioral intentions, and behaviors involving that year's Decennial Census. A few comparisons indicated that the two data streams yielded similar results, but the two methods frequently yielded different results, often strikingly so, and the results yielded by the probability samples seem likely to be the more accurate ones.

KEYWORDS: Nonprobability sampling; Relations between variables; Trends over time.

1. INTRODUCTION

Online survey data collection has become increasingly popular in recent years and is thought to have a series of advantages over traditional methods. For

JOSH PASEK is Associate Professor, Department of Communication and Media, Faculty Associate, Center for Political Studies, Institute for Social Research, and Core Faculty, Michigan Institute for Data Science at the University of Michigan, Ann Arbor MI, USA. JON A. KROSNICK is the Frederic O. Glover Professor in Humanities and Social Sciences, Professor of Communication, Political Science, and (by courtesy) Psychology at Stanford University and University Fellow at Resources for the Future, Stanford University, Stanford, CA, USA.
*Address correspondence to Josh Pasek, Department of Communication and Media, Faculty Associate, Center for Political Studies, Institute for Social Research, and Core Faculty, Michigan Institute for Data Science at the University of Michigan, Ann Arbor MI, USA; E-mail: jpasek@umich.edu.

doi: 10.1093/jssam/smz059

example, web-based surveys allow for rapid administration (Fricker and Schonlau 2002; Wright 2005) and dynamic visual presentation of stimuli to respondents (e.g., videos) in ways not possible in telephone or mail surveys (Couper 2000; Wright 2005).

Some online surveys have been conducted using probability samples of the population of interest, but many more online surveys have involved collecting data from members of large panels of individuals who were recruited using nonprobability sampling. By selecting individuals from these panels and applying post hoc adjustment techniques (e.g., weighting and matching), firms using these methods can field surveys for which respondents mirror the nation on certain demographic attributes at a lower cost than they could if samples were instead derived using probability sampling (Brick 2011).

The extent to which such samples produce accurate measurements has been the subject of considerable debate (Baker, Brick, Bates, Battaglia, Couper et al. 2013). Some scholars have argued that sample composition bias is an inevitable limitation for these studies (e.g., Langer 2018), whereas others have asserted that properly adjusted data from nonprobability samples can yield accurate assessments (e.g., Wang, Rothschild, Goel, and Gelman 2015). Most side-by-side evaluations to date have concluded that probability sample surveys yielded more accurate results—sometimes strikingly more accurate—than did nonprobability sample surveys when comparing measurements of proportions of people with specific characteristics to known population proportions (Baker, Blumberg, Michael Brick, Couper, Courtright et al. 2010; Yeager, Krosnick, Chang, Javitz, Levendusky 2011; Pasek 2016; MacInnis, Krosnick, Ho, and Cho 2018; Sohlberg, Gilljam, and Martinsson 2017; for a review, see Cornesse, Blom, Dutwin, Krosnick, de Leeuw et al. 2020). And to date, no reliable method has been identified to *a priori* eliminate the inaccuracies observed in nonprobability sample measurements without also knowing the truth (either from a parallel probability sample survey or from other records).

In an important report reviewing the growing use of nonprobability samples and their applicability, the American Association for Public Opinion Research's Task Force on Nonprobability Sampling noted that accumulated evidence indicated greater accuracy from probability samples than from nonprobability samples but also noted that the reduced accuracy of the latter might nonetheless be acceptable when pursuing some research goals. That is, nonprobability samples might sometimes be "fit for purpose" (Baker et al. 2013).

Two particular purposes are of great interest to many researchers: gauging relations between variables and studying trends over time. Even though nonprobability samples yield relatively inaccurate measurements of variable distributions, those errors may be consistent across population subgroups and over time in multiple data collections, which would allow researchers to document relations between variables and trends over time reasonably well with nonprobability samples. If so,

the value of nonprobability sampling methods rises considerably. The study reported in this article was designed to gauge the extent to which probability sample telephone surveys and nonprobability sample internet surveys yield consistent evidence about relations between variables and trends over time.

We also explored another timely but understudied issue in this arena: replicability of comparisons. During the last decade, the sciences have become increasingly interested in the replicability of findings, on the assumption that replicable findings merit more credibility. In the literature exploring the impact of mode and sampling on survey results, most studies have been one-offs, comparing a pair of data sets with one another (for an exception, Yeager et al. 2011). The data analyzed here allow for the conduct of multiple independent tests to explore robust conclusions about differences between the results obtained via different data collection methods.

1.1 Justifications for Using Nonprobability Sampling

Scholars have offered three types of rationales for using nonprobability samples to learn about populations. Psychology, for example, has traditionally employed a "generalize until proven otherwise" approach, relying on data collected from students or paid workers on Amazon's Mechanical Turk without implementing any systematic procedure for selecting potential participants from any population. And psychologists have routinely made the untested assumption that the phenomena documented in one study from a set of studies are fundamental to human nature (cf. Kam, Wilking, and Zechmeister 2007).

Other researchers have implemented matching, calibration, and/or poststratification to weight nonprobability samples so that they more closely resemble the population of interest in terms of some variables, usually demographics (Rivers 2006; Lee and Valliant 2009). Such adjustment approaches have been frequently employed to justify claims that their samples are "representative" (see Cornesse et al. 2020).

Finally, methods have been proposed to use nonprobability sample data to estimate relations of an outcome variable to various predictors and then to use probability sample data measuring those predictors to synthesize population estimates of the outcome variable (see Elliott and Valliant 2017; Sakshaug, Wiśniowski, Ruiz, and Blom 2019).

The latter two approaches are based on one of two assumptions: (1) the substantive conclusions being reached are homogeneous across sub-populations, or (2) some combination of variables available to researchers can fully eliminate differences between probability and nonprobability samples (cf. Kohler, Kreuter, and Stuart 2019). The plausibility of these assumptions is an empirical matter specific to each investigation, not an issue that can be resolved based on theory alone or based on a handful of empirical studies. That is, we can only

determine whether relations between variables are robust to sampling method by testing whether this is true. Likewise, we can only learn whether comparisons over time are equivalent in probability samples and nonprobability samples if we test that assumption in the context of a particular investigation regarding specific samples and specific variables.

Misestimations of relations between variables in a nonprobability sample will occur if two types of bias are present: (1) subgroups must be represented in the wrong proportions in the sample, and (2) the relation between the variables of interest must be different across the subgroups in the population. The presence of both of these types of bias is sufficient to compromise inferences about associations between variables. Therefore, inferences about relations between variables will be robust to differences in sampling strategy if either of these two conditions is not met, which may happen often (cf. Berrens, Bohara, Jenkins-Smith, Silva, and Weimer 2003).

In line with this reasoning, some studies comparing probability and nonprobability samples have found them to yield relatively similar relations between variables (e.g., Alvarez, Sherman, and VanBesalaere 2003; Berrens et al. 2003; Sanders, Clarke, Stewart, and Whiteley 2007; Pasek 2016; Ansolabehere and Schaffner 2014). However, other comparisons have documented strikingly different patterns of relations (e.g., Malhotra and Krosnick 2007). Clearly, more research is needed to gauge the plausibility of the assumption that associations between variables are robust to sampling.

Few studies have explored whether probability sample surveys and nonprobability sample surveys yield similar trends over time. The most compelling evidence suggesting that trends might be robust to sampling is evidence of so-called "parallel publics," whereby societal shifts in social attitudes have been shown to appear consistently across many demographic subgroups in America (cf. Page and Shapiro 1992; Wlezien 1995). Therefore, even if nonprobability samples misrepresent the prevalence of various demographic subgroups, patterns of change over time in such samples may match those for the population as a whole. However, Pasek (2016) documented notable differences between probability and nonprobability samples in terms of trends over time in substantive measurements. Again, more work is needed.

1.2 The Current Study

During a thirteen-week period in 2009 and 2010, the U.S. Census Bureau commissioned the inclusion of questions in daily probability sample telephone surveys via random-digit dialing (RDD) done by the Gallup Organization, and e-Rewards conducted weekly surveys via the internet of members of their nonprobability opt-in panel. During each interview, respondents reported their intent to complete the 2010 Decennial Census form, whether they had completed it, a variety of purported predictors of intentions and behavior, and demographic characteristics.

Thus, the data streams offer thirteen opportunities to conduct the same type of comparison to gauge replication of cross-sectional findings. This is the first ever comparison of probability and nonprobability samples affording such extensive built-in replication. Because the questionnaire included a variety of measures thought to predict inclination to complete the Census forms, the data could be used to assess the robustness of relations between variables and trends over time across sampling/modes.

2. METHODS

2.1 Data

2.1.1 Probability sample telephone data collection. Between December 3, 2009, and April 24, 2010, the Gallup Organization (2010) completed approximately one thousand telephone interviews per day via landlines and cellular telephone numbers (AAPOR RR3 = 19.4 percent). A subset of those respondents were randomly assigned to answer the Census Bureau's questions—21 percent of the sample on average per day, ranging from 180 to 775 people. On most days, between 200 and 250 respondents answered the census questions; the median number was 216 (for additional information on the telephone samples, see appendix A of the online supplementary material).[1]

2.1.2 Nonprobability sample internet data collection. To recruit members of its opt-in online panel, e-Rewards partnered with commercial companies, including airlines, video stores, booksellers, and electronics retailers. Consumers who had relationships with participating companies (e.g., members of the British Airways frequent flyer program) were invited to join the panel and complete online questionnaires regularly. Only invited individuals were eligible to join. E-Rewards regularly examined the demographic profile of its panel members and sought new partnerships with companies that catered to demographic groups that were underrepresented. In exchange for completing questionnaires, panel members were awarded points that could be redeemed for prizes.

An internet survey was fielded each week between October 27, 2009, and December 8, 2009, and between January 18, 2010, and April 19, 2010. Each week, a stratified random sample of panel members who lived in the United States was invited to complete the questionnaire. The sampling was designed to produce completed questionnaires from nine hundred people per week, resembling the nation's adult population in terms of sex, age, race, education, and region, and including at least one hundred white, one hundred black, one hundred Hispanic, and one hundred Asian-American individuals (for more information on the internet survey methodology, see appendix A of the online supplementary material).

1. No telephone interviews were conducted on March 19, 2010, or April 4, 2010.

2.1.3 Creating comparable data streams. The analyses reported here used data collected only during the weeks when data were collected by both firms. Telephone data were aggregated for each seven-day period during which each internet survey occurred. During each week, more respondents completed the internet questionnaire on days earlier in the week than on days later in the week. In case the day of the week affected the answers respondents provided, we created base weights for the telephone data for each week to match the proportions of internet respondents who completed the questionnaire on each day. For example, if 25 percent of the internet survey's respondents answered the questionnaire on the first day of a week, the respondents interviewed by telephone that day were weighted so they would constitute 25 percent of the weighted telephone sample. As a result, when base weights were applied, the proportion of interviews completed on each day during each week was the same in the telephone and internet data.

2.1.4 Additional base weighting of the telephone samples. Base weights were also constructed to adjust for known unequal probabilities of selection based on (1) the number of adults living in the household of a respondent interviewed via a landline, and (2) the number of telephone numbers that could reach the respondent. In addition, the base weights took into account whether the respondents could be reached by a landline phone only, by a cell phone only, or by a landline and a cell phone, so that the proportions of respondents in each of these groups matched the known proportions in the population.

2.1.5 Poststratification weighting. We constructed poststratification weights for each week of telephone survey data via raking, beginning with the base weights (for details, see appendix A of the online supplementary material). The procedure to build these weights was designed by a blue-ribbon panel of survey statisticians assembled by the American National Election Studies (DeBell and Krosnick 2009) and was implemented using the *anesrake* package in R (Pasek 2012). The same method was used to poststratify the internet data, though with no base weight.

2.2 Measures. Comparisons across the data streams could be done with ten demographics: sex, race/ethnicity, age, education, marital status, census region, language spoken at home, home ownership, number of persons in the household, and presence of children in the household. Comparisons could also be made with nondemographic measures in four categories: beliefs, events, behavioral intentions, and behaviors. These substantive questions asked about whether respondents trusted the confidentiality of the census, thought filling it out would take too long, thought it was important for the census to count everyone, believed that their responses did not matter,

believed that the census could help them and could harm them, and believed the census was used to identify people in the country illegally. People also reported whether they had received the census form, whether they had completed the census form (among those who had received it), and how likely it was that they would complete the census form (among those who had not already done so). Respondents were asked the first two of these questions only after the forms had been mailed, which occurred in the middle of week eight of the study (for question wordings and codings, see appendix B of the online supplementary material).[2]

A few of the questions were asked differently in the telephone and internet modes. Respondents in the internet mode were offered an explicit "don't know" response option for most questions, whereas no such option was offered during the telephone interviews (though volunteered "don't know" responses were recorded as such). For questions involving rating scales ranging from strong agreement to strong disagreement, a middle alternative, "neither agree nor disagree," was offered to the internet respondents but not the telephone respondents.

To compare the two data streams, variables were coded in various ways. For the demographics, correlations between variables and trends over time were compared across data streams coding each demographic dichotomously, identifying people in the modal category and those not in the modal category. For the regressions, each demographic was represented by a series of dummy variables. For opinion questions answered on an agree-disagree rating scale, respondents were coded dichotomously as either agreeing or disagreeing, and comparisons of trends over time examined the proportion of people who agreed divided by the proportion who agreed or disagreed.[3] For the question asking whether the census could benefit the respondent, harm the respondent, or do neither, two dummy variables were created, identifying people who thought the census could benefit them and people who thought the census could harm them.[4]

Regarding intentions to fill out and mail in the census form, we created three variables: (1) an ordinal rating scale ranging from "definitely would not complete the form" to "definitely would complete it," (2) a dummy

2. The telephone questionnaire was updated to include questions about whether the form had been received and whether they had completed it part way through week eight, whereas these questions were only asked of internet respondents starting at the beginning of week nine. Because of this difference, we dropped data on all outcome measures in week eight to maximize compatibility.

3. Respondents who chose "neither agree nor disagree" were treated as missing when we computed distributions or correlations and were given values determined by multiple imputation when we conducted regressions. Alternative coding strategies for these measures tended to reduce correspondence between the probability and nonprobability samples.

4. Telephone respondents could volunteer the answer "both" to the benefit/harm question, and those who did so were coded into both categories.

variable comparing people who said they would definitely complete the form to all others, but treating people who said they had completed the form already as missing, and (3) a dummy variable comparing respondents who said that they either definitely or probably would *not* complete the form to respondents who offered higher likelihoods of completing the form, treating people who said they had already completed the form as identical to those who reported that they would definitely complete it (for details on coding of all questions, see appendix B of the online supplementary material).

2.3 Analysis strategy. Statistical analyses were designed to address three questions:

(1) How similar are correlations between variables across the data streams?
(2) Do regressions yield similar findings across the data streams?
(3) Do the two data streams yield similar estimates of which variables changed over time and of the magnitude of those changes?

For each of these questions, we assess (1) whether there was a pattern in the probability sample telephone data, (2) whether a similar pattern emerged in the nonprobability sample internet data, and (3) whether the patterns in the two data streams were significantly different from one another.

2.3.1 Relations between variables. For every pair of variables among the ten demographic and eleven substantive measures, we computed the Pearson product moment correlation in each week in each data stream. And again for each pair of variables, we subtracted the internet correlation from the telephone correlation and computed the absolute value of the difference. Because measures of form receipt and completion were not asked before week eight (and measures asked inconsistently in week eight were dropped), there were 170 possible comparisons made for weeks one to seven, 136 potential comparisons for week eight, and 210 possible comparisons for weeks nine to thirteen, for a total of 2,376 potential comparisons across data streams of correlations between variables. After dropping 20 comparisons between variables with linear dependencies (e.g., Census completion and intention to complete the Census) and five cases where no respondents answered some pairs of response options, 2,351 bivariate comparisons were used.

We estimated the parameters of two sets of two ordinal logistic regression equations using an array of demographics and opinions about the census to predict (1) intentions to complete the census form among individuals who had not already completed it (represented as an ordinal variable ranging from "definitely will not complete the form" to "definitely will complete it"), and (2) completion

of the form using either the unweighted or weighted data.[5] Each regression equation included a dummy variable representing data stream and interactions of that dummy variable with all other predictors (see Equation 1, where y is the dependent variable, X is the array of demographic and substantive predictors, and the data stream is represented by a dummy variable indicating whether the data came from the telephone or internet samples). The parameters of the regression were estimated twice for each week, once with the telephone mode coded as the reference category, and again with the internet mode coded as the reference category, thus allowing us to obtain the results of various needed tests of statistical significance.

$$y = X + (X \times \textit{data stream}) + \textit{data stream} \tag{1}$$

2.3.2 Trends over Time

We implemented three different methods for assessing the similarity across modes of variation in substantive measures over time. These involved (1) examining how strongly correlated the weekly point estimates were between the two data streams, (2) testing whether interactions between weekly dummy variables and a data stream dummy variable produced estimates that fit the data significantly better than estimates that did not include an interaction between these terms, and (3) examining whether linear time trends predicting the weekly estimates across data streams yielded differing slopes. All of these methods allowed for mean differences between the samples in the distributions of responses. The first method also was not sensitive to whether variations within the samples were of different absolute magnitudes.

To determine whether the two data streams told similar stories about trends over time, we first calculated Pearson correlations between the weekly estimates of the proportion of respondents selecting answers in each data stream. We then estimated the parameters of a series of logistic regression equations in which dummy variables for each week and a dummy variable indicating survey data stream were used

5. Because beliefs about the census were sometimes measured slightly differently across data streams (e.g., in terms of the presence of middle response categories and expressed "don't know" options), we explored a series of different coding strategies for these measures. The results we describe in this article (where predictors were dichotomized and where middle response categories and "don't know"s were initially treated as missing and were subsequently imputed using multiple imputation via chained equations) produced results that were most similar in the two data streams. Alternative coding strategies included coding the variables incrementally in the range from zero to one, as well as omitting missing data rather than imputing. Results from these alternative strategies increased the number of differences observed between data streams. Cases that did not answer an outcome measure question were dropped from regressions. Imputation was done using the *MICE* software for R (van Buuren and Groothuis-Oudshoorn 2011). After computing five imputations for each data stream within each week, the imputed data were merged to yield five multiply imputed data sets for each of the thirteen weeks.

to predict the values of each target variable. By comparing two equations—one in which weekly dummy variables were interacted with data stream (Equation 2) and one where they were not (Equation 3)—we could assess whether there were statistically significant differences in trends across the two data streams. Variations between the trends in the different data streams were compared using Wald tests.

$$y = \beta_1 \text{data stream} + \sum_{i=2}^{12} \left(\beta_i Week_i + \zeta_i (Week_i \times data\ \ stream) \right) \quad (2)$$

$$y = \beta_1 data\ \ stream + \sum_{i=2}^{12} (\beta_i Week_i) \quad (3)$$

2.4 Presentation of Results. We present all analyses in this article using two distinct approaches. One strategy involved running analyses using the raw data gathered from each data stream. The other compared the two data streams after data were adjusted to yield comparable distributions of respondents within weeks and all weighting methods were employed. Although a variety of other strategies might be justifiable, we include each of these two for a few related reasons. First, completely unweighted analyses reveal the variability of the underlying respondent selection procedures over time. All weighting strategies overstate the variability associated with sampling by assigning different weights to different cases. Completely unweighted analyses also avoid the potential for known but difficult to correct biases in inferences associated with the use of survey weights (see Gelman 2007). Presenting unweighted results helps to mitigate this concern. In contrast, the presentation of fully weighted and date-matched results is important for ensuring that differences between the data streams are not simply a product of differences in the types of people completing the study or the timing at which they did so. These results allow us to understand how the two data streams differ in the conclusions they support about the target population. Alternative specifications yielded similar results; these are not presented for the sake of parsimony.

Because of the potential sensitivity of weighted data to variance inflation, analyses for the current project were generated using a series of parametric bootstraps. Cases from each data stream within each week were resampled five hundred times with replacement. The p values for comparisons across streams were assessed by pairing these resampled data sets and were assessed as twice the proportion of resamples for which the parameter in the generally smaller data stream was larger than the parameter in the generally larger one (to yield a two-tailed test). In the regression analyses, where multiple imputation was used to address missing data and differences in measurement across data streams, bootstrapped cases were resampled from a data set that included all cases imputed across five unique multiple imputations. The number of resampled cases was set to match the number of cases present before imputation.

Finally, our interest in this article is in understanding the extent to which data from the nonprobability sample internet data will yield conclusions that match those of the probability sample telephone data. This leads us to pay particular attention to analyses where the telephone results tell a statistically significant story about relations between variables or trends over time and thus where researchers would typically reject the null hypothesis if they were using the traditional sample. Focusing on these analyses, we examine whether the results from the data streams would lead to substantively different conclusions about the magnitude or presence of relations for each type of inference we examine.

3. RESULTS

3.1 Correlations Between Variables

The conclusion a researcher would reach when examining a correlation between a pair of variables in terms of (1) direction (positive or negative), (2) magnitude, and (3) statistical significance would sometimes be the same using the two data streams, would sometimes be moderately different, and would sometimes be strikingly different. The x axis of figure 1 shows the magnitude of each correlation in the probability sample telephone data, and the y axis shows the absolute value of the difference between that correlation and the comparable correlation observed in the nonprobability sample internet data. When using base weights only and all pairs of variables (2351 pairs), only 57.8 percent of the correlations in the two data streams yielded what we call "same story, same magnitude," meaning that the correlations are in the same direction, are either both statistically significant or are both nonsignificant, and do not differ significantly from one another. That 57.8 percent can be decomposed into two groups: 35.5 percent of the pairs of variables were uncorrelated in both data streams (small gray circles in figure 1), and 22.3 percent of the pairs of variables were significantly correlated with one another at the same magnitude (small black circles in figure 1) and yielded the same indication about statistical significance (i.e. were both significant).

Thus, for nearly half of the pairs of variables, the two data streams yielded different substantive conclusions. For one third of the variable pairs, the two data streams would lead a researcher to reach different conclusions about whether the variables are correlated significantly in a particular direction, 13.9 percent in which the two correlations were significantly different from one another (gray diamonds in figure 1) and 19.8 percent in which the two correlations were not significantly different from one another (black-bordered diamonds in figure 1). For another 7.0 percent of correlations, a researcher would reach the same conclusions from the two correlations about direction and significance, but the two correlations are significantly different

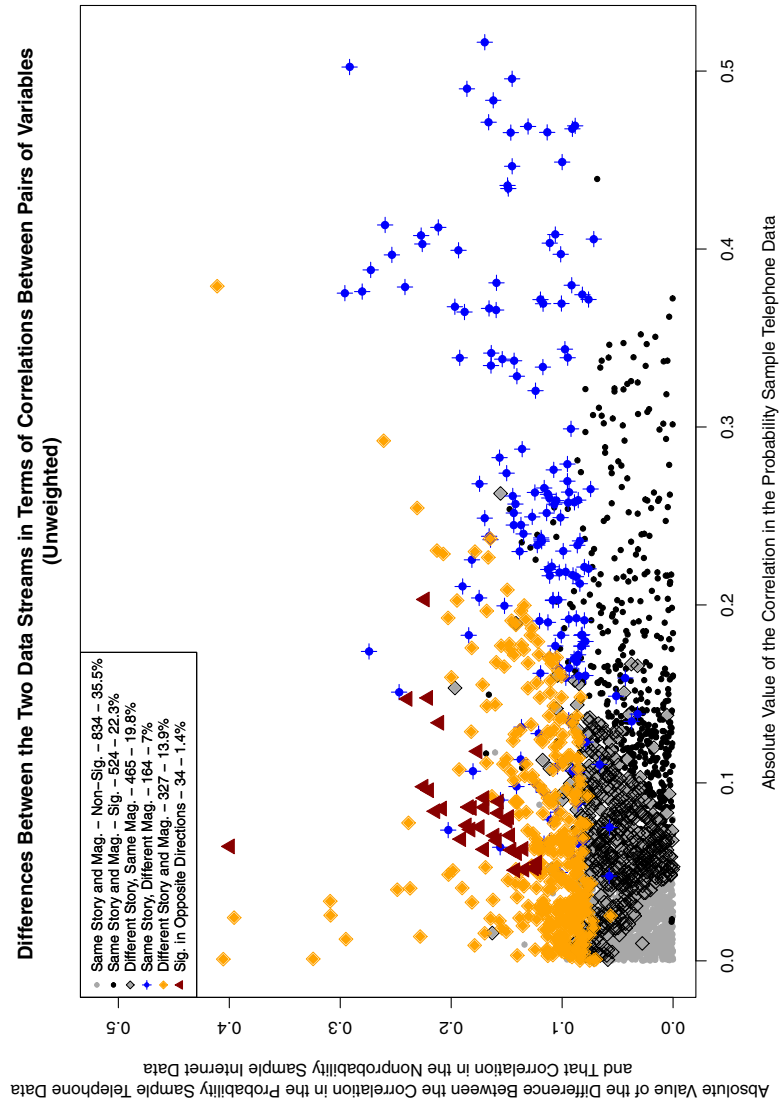

Figure 1. Differences Between the Two Data Streams in Terms of Correlations Between Pairs of Variables (Unweighted).

from one another in magnitude (stars in figure 1). Finally, for 1.4 percent of the correlations, the two data streams yielded correlations that were each individually significant with opposite signs and that were significantly different from one another (triangles in figure 1). Thus, conclusions from correlations would routinely be quite different depending on which data stream was used.

After poststratification and matching the numbers of interviews per day in the two data streams, the correlations between pairs of variables were even more different from one another (see figure 2). The number of correlations telling the same story about direction, magnitude, and statistical significance dropped to 49.2 percent, meaning that the two data streams told different stories about correlations more than half of the time. Thus, the two data streams yielded more different results after weighting and date matching than before.

3.2 Regression Coefficients

As with the correlations, the conclusion a researcher would reach when examining regression coefficients in terms of (1) direction (positive or negative), (2) magnitude, and (3) statistical significance would sometimes be the same using the two data streams, would sometimes be moderately different, and would sometimes be strikingly different. The x axis of figures 3a-d show the absolute value of the magnitude of each regression coefficient in the probability sample telephone data, and the y axis shows the absolute value of the difference between that coefficient in the probability sample telephone data and the comparable coefficient observed in the nonprobability sample internet data.

When predicting reported likelihood of completing the census form without weights or date matching, 70.5 percent of the coefficients in the two data streams yielded what we call "same story, same magnitude," meaning that the coefficients are in the same direction, are either both statistically significant, or are both nonsignificant and do not differ significantly from one another. That 70.5 percent can be decomposed into two groups: 49.6 percent of the coefficients are zero in both data streams (the gray circles in figure 3a), and 20.9 percent of the coefficients are significantly different from zero in both data streams (the black circles in figure 3a).

Thus, for one quarter of the regression coefficients, the two data streams yielded different substantive conclusions. For 24.4 percent of the coefficients, the two data streams would lead a researcher to reach different conclusions about whether the variable is a significant predictor in a particular direction, 8.1 percent in which the two coefficients were significantly different from one another (diamonds in figure 3a), and 16.2 percent in which the two coefficients were not significantly different from one another (black-bordered diamonds in figure 3a). For another 4.3 percent of coefficients, a

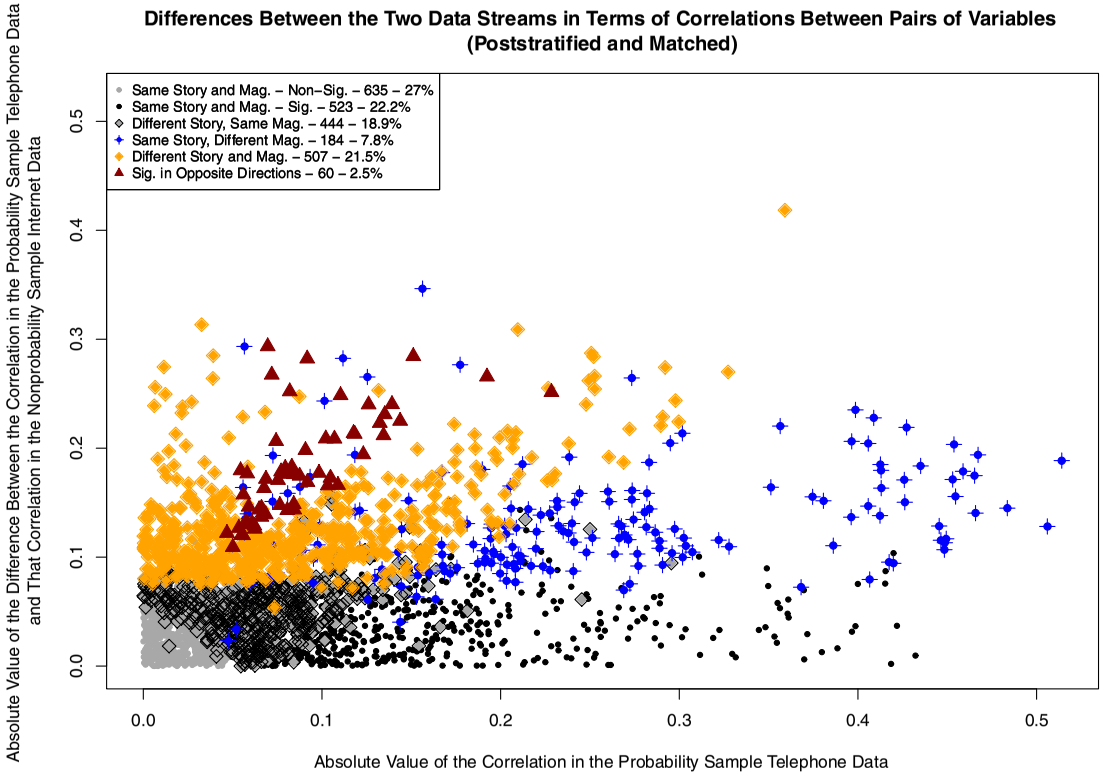

Figure 2. Differences Between the Two Data Streams in Terms of Correlations Between Pairs of Variables (Poststratified and Matched).

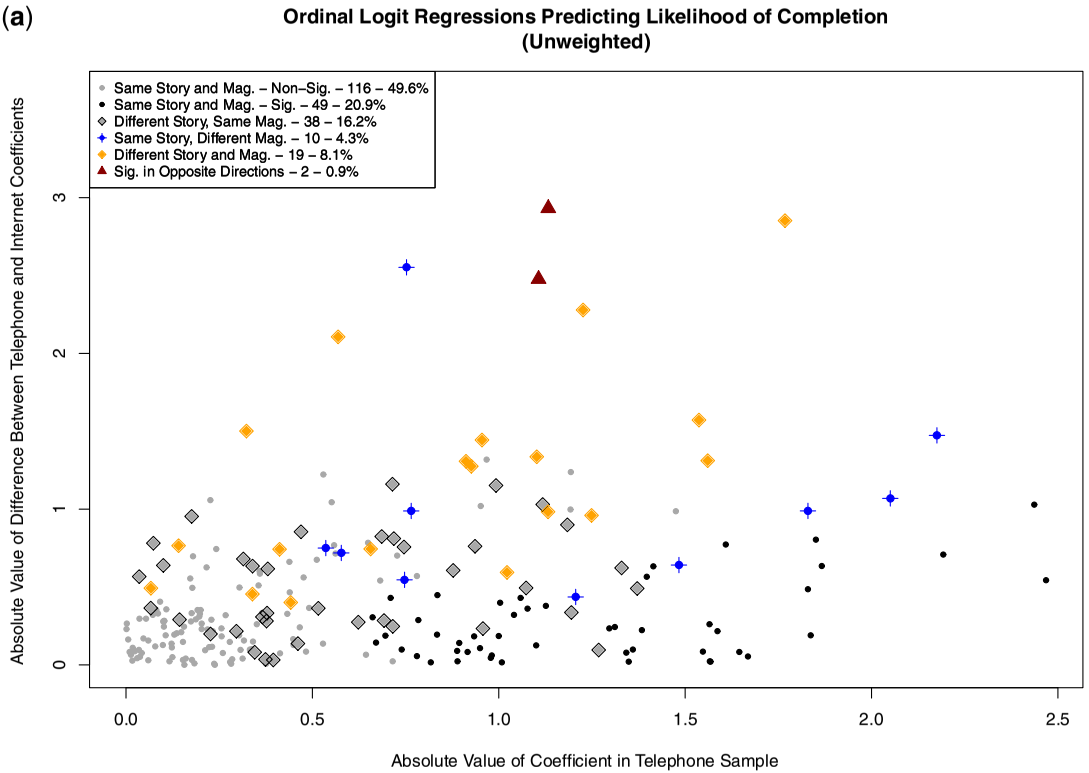

Figure 3. (A) Ordinal Logit Regressions Predicting Likelihood of Completion (Unweighted); (B) Ordinal Logit Regressions Predicting Likelihood of Completion (Poststratified and Matched); (C) Regressions Predicting Reported Completion (Unweighted); (D) Regressions Predicting Reported Completion (Poststratified and Matched).

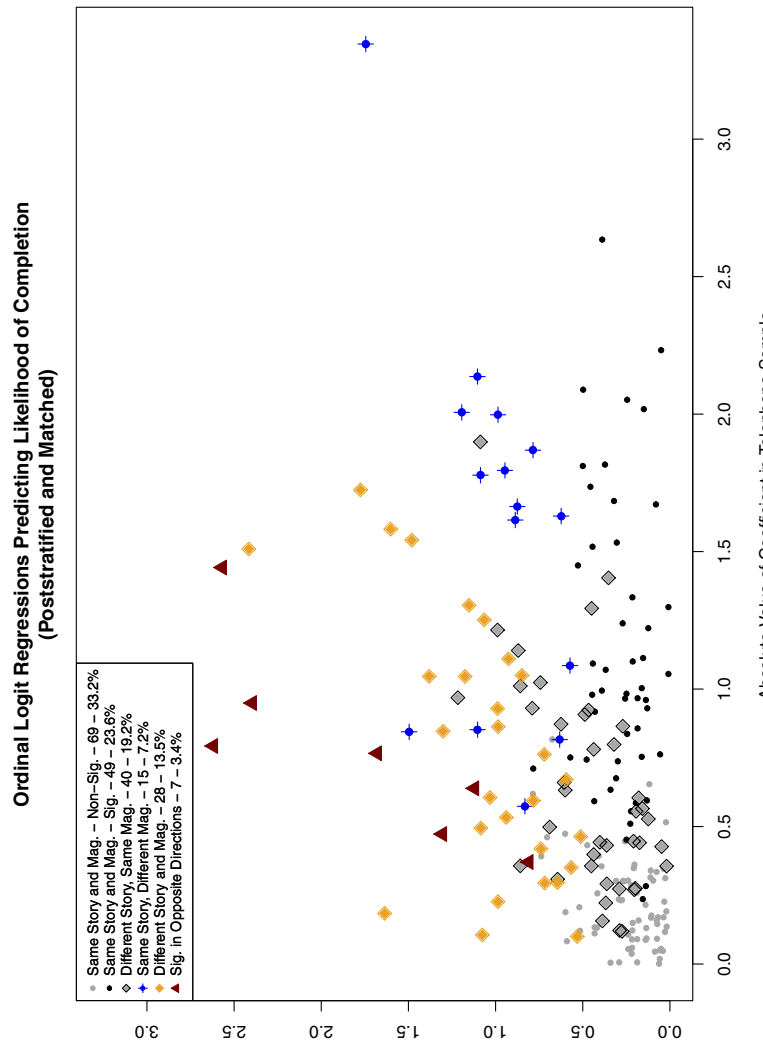

Figure 3. Continued

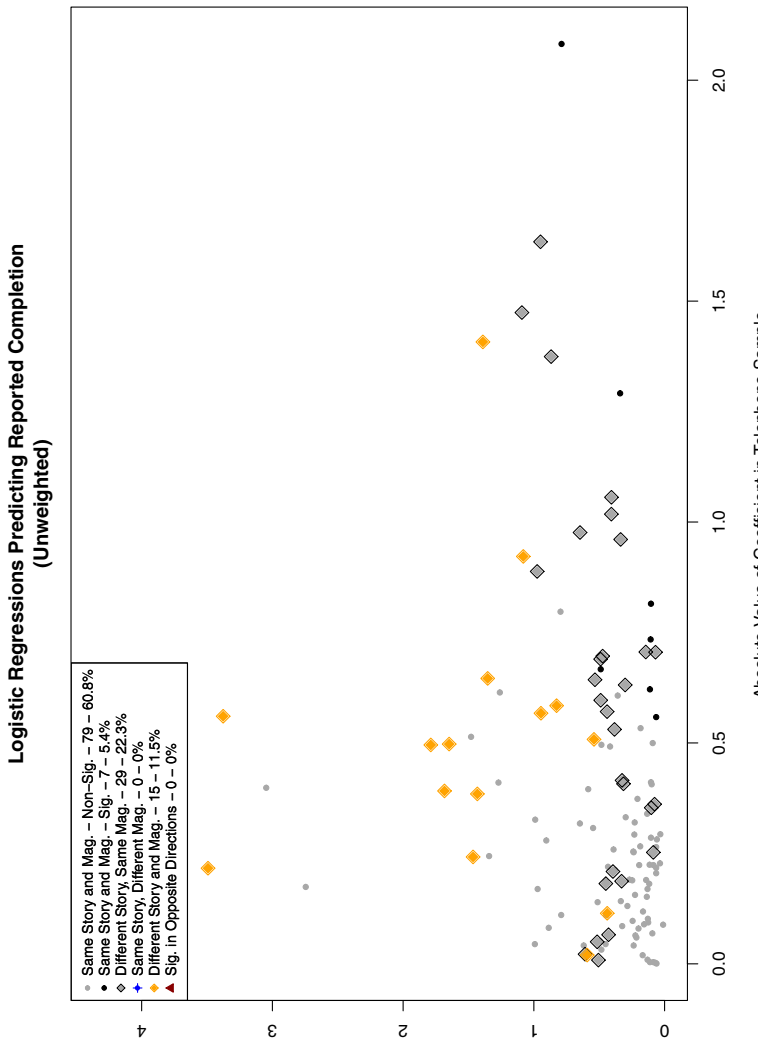

Figure 3. Continued

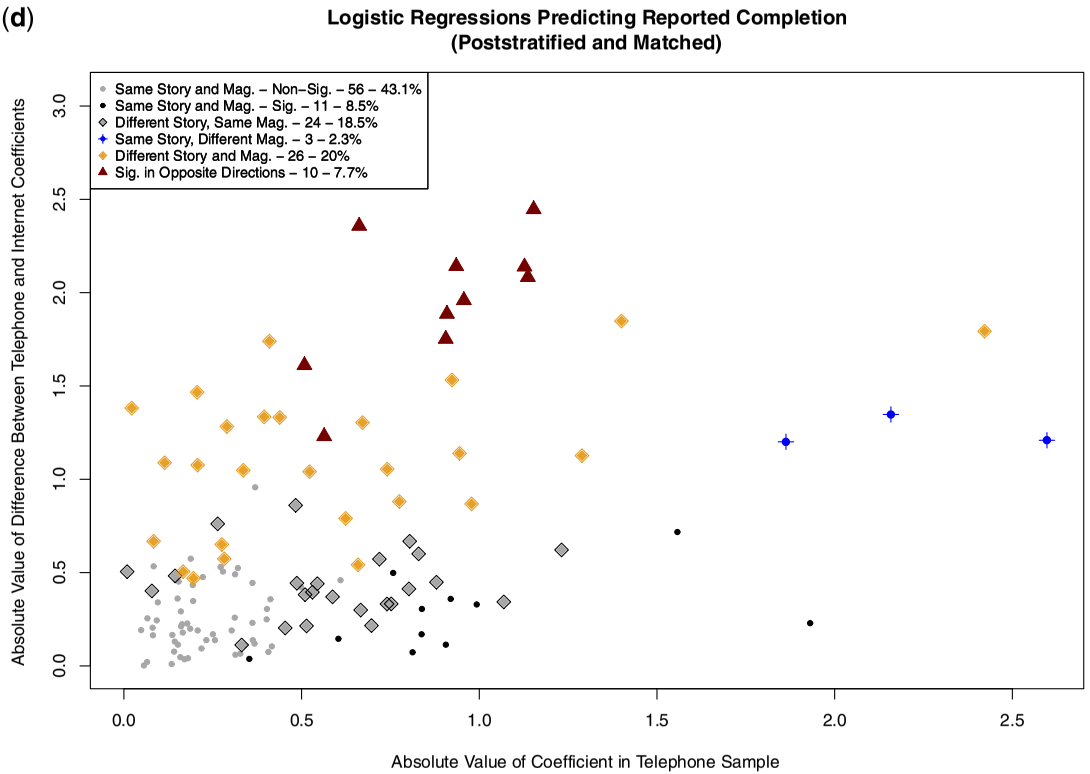

Figure 3. Continued

researcher would reach the same conclusions from the two coefficients about direction and significance, but the two coefficients are significantly different from one another in magnitude (stars in figure 3a). Finally, for two of the coefficients (0.9 percent), the two data streams each yielded significant coefficients with opposite signs that were significantly different from one another (triangles in figure 3a). Thus, about one quarter of the conclusions reached based on regression coefficients would be different depending on which data stream was used.

After poststratification and matching the numbers of interviews per day in the two data streams, the regression coefficients were even more different from one another (see figure 3b). The number of coefficients telling the same story about the direction, magnitude, and statistical significance dropped to 56.7 percent, meaning that the number of instances in which the two data streams told different stories about coefficients increased to 43.3 percent. Thus, the two data streams yielded more different results after weighting and date matching than before.

When predicting completion of the census form without weights or date matching, 66.2 percent of the coefficients told the same story about the direction, statistical significance, and magnitude of the relation (gray and black circles in figure 3c). And that figure dropped to 51.5 percent after poststratification and date matching (figure 3d).

Taken together, these results show frequent meaningful discrepancies between the conclusions a researcher would reach using the two data streams.

3.3 Trends over Time

To assess the degree to which the two data streams told the same story about trends over time, we calculated change over time in each measure between each pair of waves in which the measure was asked, and we compared that change across the two data streams. If the data streams yielded equivalent indications of trends over time, these pairwise differences should have been the same in the two data streams.

When examining unweighted data without date matching, the data streams told the same story about the direction, magnitude, and significance of the change between waves for 73.5 percent of the wave pairs (shown by the gray circles and black circles in figure 4). That is, a researcher would reach different conclusions about changes over time using the two data streams for 26.5 percent of the wave pairs. When using poststratification and date matching the proportion of interviews completed per day, the number of wave pairs about which the researcher would reach the same conclusion from both data streams is 67.3 percent (see figure 5).

Thus, more than 30 percent of wave pairs yield different conclusions in the different data streams.

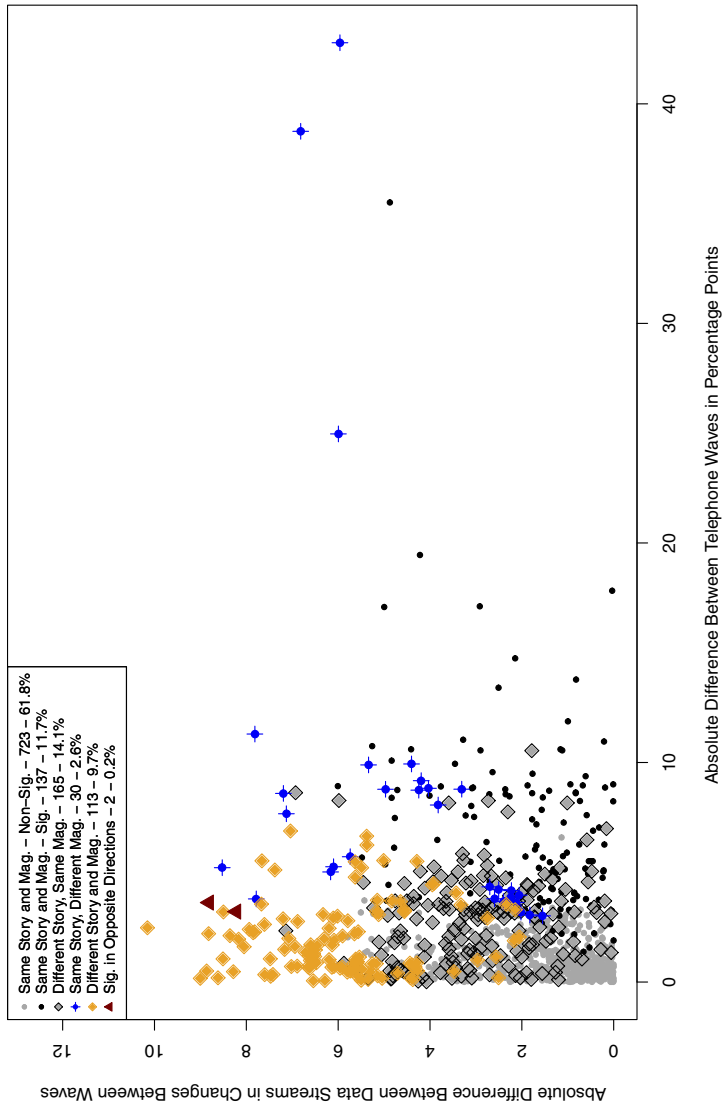

Figure 4. Differences Between Pairs of Waves Across Data Streams (Unweighted).

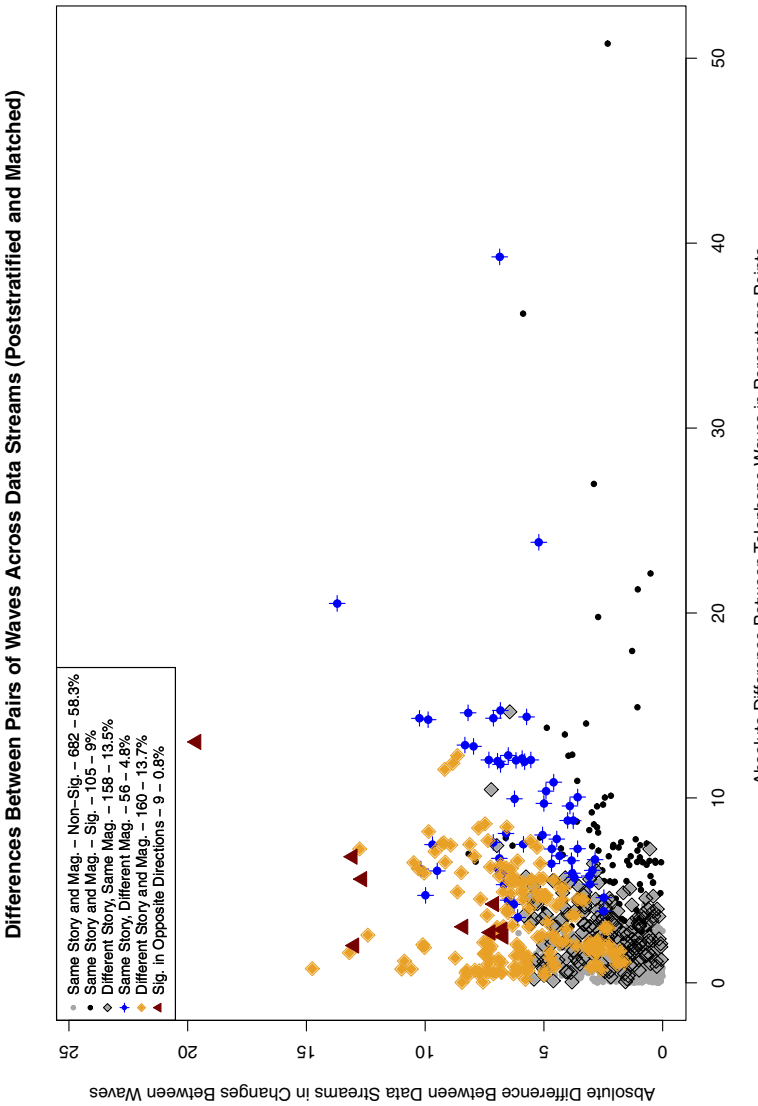

Figure 5. Differences Between Pairs of Waves Across Data Streams (Poststratified and Matched).

4. DISCUSSION

This investigation revealed frequent and sometimes sizable differences between probability sample telephone data and nonprobability sample internet data in terms of correlations among variables, predictors of behavioral intention and behavior, and changes over time. These differences were often replicated across survey weeks and were robust across a variety of poststratification weighting strategies. Clearly, nonprobability sample internet data did not consistently yield the same results as probability sample telephone data.

For researchers deciding whether to conduct probability sample telephone surveys or nonprobability sample internet surveys, the results presented here suggest that this choice is indeed likely to be consequential if studying relations between variables or trends over time. Just as in past studies examining the accuracy of distributions of responses to items, the evidence here shows some instances of correspondence between data streams but enough instances of mismatches to suggest strong caution before presuming that nonprobability sample internet data will yield the same conclusions as probability sample telephone data.

The present results regarding differences between data streams in terms of regression coefficients stand in contrast to the findings of some prior studies that documented closer matches (Alvarez et al. 2003; Ansolabehere and Schaffner 2014; Berrens et al. 2003; Sanders et al. 2007). One likely explanation for this difference stems from the political nature of most survey comparisons to date. Because the regressions in these studies controlled for partisanship and ideology, which were closely related to the outcomes of interest, these earlier studies may have used questions for which matches between data streams were particularly likely. In contrast, the predictors of the outcomes of interest here yielded more mismatches in results.

4.1 Limitations

Like many other attempts to understand how conclusions from probability and nonprobability samples relate, this study is limited by the fact that the two data streams were not collected for the primary purpose of comparing nonprobability and probability sampling. For example, differences in the response options offered for some questions across data streams could have extenuated differences between them. Similarly, the discrepancies we observed could be attributable, in part, to the differences in respondent behavior between the survey modes. This means that it is not possible for the present study to distinguish mode differences from differences due to probability versus nonprobability sampling, as well as due to question wording.

An ideal comparison to understanding the efficacy of inferences from nonprobability samples versus probability samples would be one that juxtaposes multiple samples that employ identical modes of data collection and questions

but that differed only in terms of sampling strategy. Were this the case here, we would be able to identify the reasons that conclusions often differed across data streams. Instead, we can only conclude that the differences between these two strategies for data collection are consequential. We cannot report a singular reason why this was the case. That said, other data suggest that the differences between probability sample telephone data and probability sample internet data are usually very small (e.g., Yeager et al. 2011). And much evidence shows sizable differences between the results obtained from people who volunteer to complete questionnaires online versus probability samples who complete online questionnaires. This suggests that the differences between data streams observed here might be more due to sampling than to the mode of data collection.

The current study employed a specific set of weighting methods using demographics. It is possible that a more comprehensive set of weighing variables or alternative weighting methods could improve the correspondences observed.

5. CONCLUSION

The results reported here showed that probability sample RDD telephone surveys and nonprobability sample internet surveys did not consistently support the same conclusions. When attempting to understand opinions and behaviors relevant to the 2010 Decennial Census, data collection method altered conclusions that researchers would reach about relations between variables, predictors of intent to complete the census form and its actual completion, and over-time trends. Therefore, it does not seem reasonable to assume that inferences about relations between variables or trends over time will be robust to mode and sampling differences. Future research should continue to assess the objective accuracy of both sets of methods and explore whether there are conditions under which these data collection methods yield more compatible results.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Alvarez, R. M., R. P. Sherman, and C. VanBeselaere (2003), "Subject Acquisition for Web-Based Surveys," *Political Analysis*, 11, 23–43.

Ansolabehere, S. D., and B. F. Schaffner (2014), "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison," *Political Analysis*, 22, 285–303.

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau (2013), "Summary Report of the AAPOR Task Force on Nonprobability Sampling," *Journal of Survey Statistics and Methodology*, 1, 90–143.

Baker, R., S. J. Blumberg, J. Michael Brick, M. P. Couper, M. Courtright, J. Michael Dennis, and D. A. Dillman (2010), "AAPOR Report on Online Panels," *Public Opinion Quarterly*, 74, 711–781.

Berrens, R. P., A. K. Bohara, H. Jenkins-Smith, C. Silva, and D. L. Weimer (2003), "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples," *Political Analysis*, 11, 1–22.

Brick, J. M. (2011), "The Future of Survey Sampling," *Public Opinion Quarterly*, 75, 872–888.

Cornesse, C., A. G. Blom, D. Dutwin, J. A. Krosnick, E. D. de Leeuw S. Legleye, J. Pasek, et al. (2020), "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research," *Journal of Survey Statistics and Methodology*, 8, 4–36.

Couper, M. P. (2000), "Review: Web Surveys: A Review of Issues and Approaches," *Public Opinion Quarterly*, 64, 464–494.

DeBell, M., and J. A. Krosnick (2009), "Computing Weights for American National Election Study Survey Data," nes012427. Ann Arbor, MI, Palo Alto, CA: ANES Technical Report Series.

Elliott, M. R., and R. Valliant (2017), "Inference for Nonprobability Samples," *Statistical Science*, 32, 249–264.

Fricker, R. D. Jr., and M. Schonlau (2002), "Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature," *Field Methods*, 14, 347–367.

Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, 22, 153–164.

Kam, C. D., J. R. Wilking, and E. J. Zechmeister (2007), "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research," *Political Behavior*, 29, 415–440.

Kohler, U., F. Kreuter, and E. A. Stuart (2019), "Nonprobability Sampling and Causal Analysis," *Annual Review of Statistics and Its Application*, 6, 149–172.

Langer, G. (2018), "The Importance of Probability-Based Sampling Methods for Drawing Valid Inferences," in *The Palgrave Handbook of Survey Research (Vol.* 74), eds. Vannette, D. L. and Krosnick J. A., 7-12, Palgrave Macmillan.

Lee, S., and R. Valliant (2009), "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment," *Sociological Methods & Research*, 37, 319–343.

MacInnis, B., J. A. Krosnick, A. S. Ho, and M.-J. Cho (2018), "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension," *Public Opinion Quarterly*, 82, 707–744.

Malhotra, N., and J. A. Krosnick (2007), "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples," *Political Analysis*, 15, 286–323.

Page, B. I., and R. Y. Shapiro (1992), *The Rational Public*. Chicago: University of Chicago Press.

Pasek, J. (2012), "Anesrake: ANES Raking Implementation." In *Comprehensive R Archive Network*, CRAN.

———. (2016), "When Will Nonprobability Surveys Mirror Probability Surveys? considering Types of Inference and Weighting Strategies as Criteria for Correspondence," *International Journal of Public Opinion Research*, 28, 269–291.

Rivers, D. (2006), "Sample Matching: Representative Sampling from Internet Panels," in *Polimetrix White Paper Series*, Palo Alto, CA: YouGovPolimetrix.

Sakshaug, J. W., A. Wiśniowski, D. A. P. Ruiz, and A. G. Blom (2019), "Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach," *Journal of Official Statistics*, 35, 653–681.

Sanders, D., H. D. Clarke, M. C. Stewart, and P. Whiteley (2007), "Does Mode Matter for Modeling Political Choice? Evidence from the 2005 British Election Study," *Political Analysis*, 15, 257–285.

Sohlberg, J., M. Gilljam, and J. Martinsson (2017), "Determinants of Polling Accuracy: The Effect of Opt-in Internet Surveys," *Journal of Elections, Public Opinion and Parties*, 27, 433–447.

van Buuren, S., and K. Groothuis-Oudshoorn (2011), "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45(3).

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015), "Forecasting Elections with Non-Representative Polls," *International Journal of Forecasting*, 31, 980–991.

Wlezien, C. (1995), "The Public as Thermostat: Dynamics of Preferences for Spending," *American Journal of Political Science*, 39, 981–1000.

Wright, K. B. (2005), "Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services," *Journal of Computer-Mediated Communication*, 10, doi: 10.1111/j.1083-6101.2005.tb00259.x.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang (2011), "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Nonprobability Samples," *Public Opinion Quarterly*, 75, 709–747.