Less Trust, Moore Verification: Determining the Accuracy of Third-Party Data through an Innovative Use of Attention Checks

Nathan Seltzer, University of Wisconsin-Madison
nseltzer@wisc.edu

SUMMARY

In the days following the publication of a *Washington Post* article that detailed allegations of sexual abuse against Roy Moore, Emerson College Polling released an election poll of Alabama voters that showed Moore maintaining a 10-point lead over his opponent Doug Jones. The Emerson College Poll was conducted using survey data administered by landline phone and over the internet. In this working paper, I analyze raw data from this poll and find irregularities in the internet sample that might suggest that the respondents were not properly sampled by the data vendor that administered the survey, Opinion Access Corp. Specifically, a substantial number of respondents in the internet sample were unable to accurately match their county of residence to their US congressional district when presented with a map of Alabama which displayed both (Figure 2). The misclassification rate for the internet sample was 36% (117 respondents out of 324). Such a high misclassification rate might indicate that some of the respondents in the sample did not reside in Alabama.

Although the aim of this paper is not to predict the outcome of an electoral contest, the removal of this poll from aggregate polling averages might indicate a tighter Alabama senate race than previously understood. Emerson College Polling released an additional poll that surveyed support for Roy Moore and Doug Jones in the Alabama senate race on November 28, 2017 that similarly relied on respondents acquired through Opinion Access Corp. If the same irregularities observed in the November 13, 2017 poll are present in the more recent poll, then political observers should interpret the results with the understanding that a substantial number of respondents interviewed might be invalidly included.

As researchers increasingly rely on internet data vendors to acquire respondents for polls and surveys, I argue for the necessity of proactively verifying the accuracy of third-party data. With the use of survey “attention checks,” researchers can determine whether data vendors have provided samples that match their requested sampling frame and gain confidence in the validity of their results.

Less Trust, Moore Verification: Determining the Accuracy of Third-Party Data through an Innovative Use of Attention Checks

Nathan Seltzer, University of Wisconsin-Madison
nseltzer@wisc.edu

INTRODUCTION

Survey researchers increasingly administer online surveys to replace or augment in-person or phone interviews (Hays and Kapteyn 2015). This practice often entails contracting out sampling and survey administering to third-party data vendors – frequently for-profit marketing or polling firms – that draw respondents from ongoing internet-based panels. A substantial body of research in the field of survey methodology has assessed the proper and improper use of non-probability sampling as well as weighting procedures used to correct for sample unrepresentativeness (Tourangeau, Conrad, and Couper 2013).

In this note, I discuss a practical, yet critical, concern regarding the use of survey data provided by internet panel vendors or any third-party data vendor. Specifically, I argue that researchers should proactively verify that third-party survey data is accurately sampled before considering it for analysis. Researchers often fail to do this accuracy check because they expect that third-party data vendors have pre-screened the sample based on their requested sampling frame criteria. Under the expectation that the provided sample has been accurately pre-screened, researchers commonly omit demographic items from questionnaires that could otherwise be used to verify whether the sample matched the requested sampling frame criteria. Importantly, researchers need to take steps to avoid the use of inaccurate samples that yield erroneous conclusions.

To illustrate the gravity of this concern, I present findings from a simple analysis of raw data from a political opinion poll administered on behalf of an academic-based polling center by a third-party internet panel vendor. This poll has received sustained national press because of its startling findings and the substantial national attention placed on the electoral contest for which it measured public opinion. By assessing valid response choices to overlapping geographic variables, I identified irregularities in the dataset that suggest that the sample included respondents who were not within the researchers' intended sampling frame. As a means of proactively verifying the accuracy of survey data provided by third-party data vendors, I propose an innovative use of survey "attention checks."

ATTENTION CHECKS FOR DATA VENDORS

Attention checks are widely used by survey researchers to assess data quality by examining whether respondents have provided sufficient attention to survey instructions while answering questionnaires. These often take the form of "trap" questions that should be correctly answered if the respondent has been cognitively engaged with the instructions of the survey, but incorrectly answered if the respondent is not paying attention or exerting minimal effort to complete the survey (Anduiza and Galais 2017).

While the conventional use of attention checks attempts to ascertain valid item responses by survey respondents, I propose that researchers use attention checks to verify whether third-party data vendors have provided samples that match sampling frame criteria requested by the researcher. I define *data vendor attention checks* as demographic questionnaire items with response categories that extend beyond the range of the intended sampling frame's demographic characteristics. Since researchers have *a priori* expectations of who should be in their sample, the distribution of responses to data vendor attention check items should correspond to their expectations. Any substantive divergence in the distribution of demographic variables used to define the sampling frame would indicate that the sample includes respondents who were not intended to be in the sampling frame. If this is the case, then the data vendor has not paid sufficient attention to providing an accurate sample to the researcher.

In the example that follows, I illustrate how researchers can repurpose pre-existing questionnaire items as attention checks to verify the accuracy of a sample provided by a third-party internet panel vendor.

AN EXAMPLE: POLLING THE SPECIAL ALABAMA SENATE RACE

Emerson College Polling, an academic-affiliated polling center at Emerson College, fielded a political survey in November 2017 that gauged support for senate candidates Roy Moore (R) and Doug Jones (D) by registered and likely Alabama voters in the special Alabama senate race to be held on December 12th, 2017. The findings of the poll received substantial state and national press given the timing of its release, which was conducted in the days following the *Washington Post* publishing an article that detailed a series of sexual abuse allegations against Roy Moore. Accordingly, the Emerson poll provided the first scientific survey of Alabama voters' candidate preferences in the wake of this major campaign shakeup. The results of the poll suggested that in a two-candidate race, Roy Moore held 55% of the vote while Doug Jones held 45% of the vote (margin of error +/- 3.9%).

Emerson College Polling is a charter member of the American Association for Public Opinion Research's Transparency Initiative (AAPOR TI) which encourages the full disclosure of data and methodology of public opinion surveys to the public as a means of transparency and replication. Emerson College Polling releases raw datasets and disclosure forms for all their publicly-released surveys. As noted on their website, "These resources will be publicly available to students, teachers, researchers, and practitioners. As an academic institution focused on advancing the understanding of public opinion research, we invite all these groups to use, study and critique our methods and results in any way they wish and to share their findings with us in a collaborative manner" (Emerson College Polling 2017).

In my analysis, I examined the Emerson College Poll from November 13, 2017. According to the press release and the methodology disclosure form, the sample was acquired through two modes of administration: interactive voice response (IVR) of landline numbers with a voter phone number file supplied by Aristotle, LLC, and an online panel survey supplied by Opinion Access Corp., LLC. The stated sampling frame consisted of registered and likely voters in the state of

Alabama. In the publicly released dataset, the IVR sample consisted of 628 respondents and the internet panel consisted of 324 respondents.[1]

ATTENTION CHECKS THROUGH *A PRIORI* EXPECTATIONS OF VARIABLE DISTRIBUTIONS

To verify whether the internet sample of the November 13, 2017 Emerson College Poll was comprised of valid Alabama respondents, I examined the joint frequency distribution of two overlapping geographic variables in the dataset: county of residence ("county") and US congressional district ("USC District"). In the internet sample, respondents were asked to indicate their county of residence from a list of all 67 Alabama counties and identify their congressional district from a map that displayed all 7 Alabama congressional districts (Figure 1). The IVR landline phone sample includes variables for county and congressional district, although the publicly-released methodology and dataset leave unclear how these questions were asked in this mode.

Examining the univariate frequency distributions of these variables would not sufficiently verify the accuracy of the sampling frame because all response categories indicate a geographic location within Alabama. Instead, I examine the joint frequency distribution of both the county and congressional district variables which reveals whether respondents indicated illogical county-district pairs.

Alabama counties are nested within congressional districts, although there are several counties that overlap with two or three congressional districts (Figure 1). As a result, we should expect that congressional districts are non-randomly distributed within counties. The *a priori* expectation of the joint distribution would be that most counties should only have respondents in one congressional district. Additionally, we should expect that respondents *correctly* matched their county and corresponding congressional district – there should be no ambiguity, with exception of the possibility of minimal respondent error.

In Figure 2, I graph the joint frequency distribution of respondents by their counties and congressional districts for both the internet sample (N=628) and the IVR phone sample (N=317). The rows of the graph correspond to county of residence while the columns correspond to the respondents' specified congressional districts. The dark blue boxes represent clusters of one or more respondents, while the light grey boxes represent no respondents. Importantly, the red x-marks indicates valid responses; all other cells in the heat map represent illogical and invalid county-district pairs.

In this figure, the IVR sample matches our *a priori* expectation for how congressional districts should be distributed within counties – most counties only have respondents in one congressional

[1] Although the dataset has 628 observations, the press release and disclosure form state that only 317 observations were used to estimate the polling results. For my purposes, I use all 628 IVR sample observations in the raw dataset. The purpose of this analysis is not to replicate the findings of the poll, but to verify the accuracy of the sample.

district. Moreover, all respondents in the IVR sample correctly matched counties and congressional districts.

For the internet sample, however, there are a substantial number of respondents who incorrectly matched their counties and congressional districts. 117 out of the 324 respondents (36.1%) were unable to accurately match their county of residence to their US congressional district when presented with a map of Alabama which displayed both. In Autauga county, for instance, which is in central Alabama and District #2, none of the respondents from the internet sample selected District #2. Instead, they indicated that their congressional district was either District #1, District #3, District #4, or District #7, all of which are incorrect.

It is unclear why respondents in the internet sample failed to correctly match their congressional districts to their county of residences. In the online questionnaire, respondents were provided a map that had congressional districts transposed over county boundaries (i.e. Figure 1 in this paper), and were then asked to indicate their congressional district. This should have been a simple task for respondents if they had knowledge of where they lived within their state of residency. To be sure, it is possible that the divergence in the joint distributions shown in the internet and IVR phone samples might have a practical explanation that is not easily inferred from the publicly-released survey methodology. But when the IVR error rate is 0% while the internet error rate is 36.1%, such an explanation seems implausible.

Overall, the findings presented here indicate the possibility that a sizeable portion of the respondents from the internet sample were not within the intended sampling frame. If this is the case, then the inferential results of the November 13, 2017 Emerson College Poll on the Alabama senate race are based on a sample that cannot be entirely verified as registered and likely voters living in Alabama. It is likely that the internet panel respondents acquired through Opinion Access Corp. were less accurately sampled than the respondents acquired through a voter file list and contacted on landline phones.

DISCUSSION/CONCLUSION

Third-party internet panel vendors provide a cost-effective and time-efficient option for conducting survey research. The findings presented in this note emphasize the importance of verifying the quality and accuracy of survey data from internet panel vendors before disseminating the findings to a broader audience. Ultimately, it is the researcher's responsibility to determine the fidelity of the data they use in their analysis.

Although the aim of this paper is not to predict the outcome of an electoral contest, the removal of this poll from aggregate polling averages might indicate a tighter Alabama senate race than previously understood. Emerson College Polling released an additional poll that surveyed support for Roy Moore and Doug Jones in the Alabama senate race on November 28, 2017 that similarly relied on respondents acquired through Opinion Access Corp. If the same irregularities observed in the November 13, 2017 poll are present in the more recent poll, then political

observers should interpret the results with the understanding that a substantial number of respondents interviewed might be invalidly included.[2]

This analysis also demonstrates the utility of using data vendor attention checks. Data vendors often have aims and motives that do not align with academic researchers, and therefore researchers should by default be skeptical of the accuracy of third party data. Besides asking data vendors to provide their methodology, researchers must take it upon themselves to create accuracy checks which they can use to determine whether the data vendor properly administered the survey.

[2] The publicly-released dataset for this poll does not include the county of residence variable for the internet sample, so I am unable to perform a data vendor "attention check" analysis.

REFERENCES

Anduiza, E., & Galais, C. (2016). Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. International Journal of Public Opinion Research.

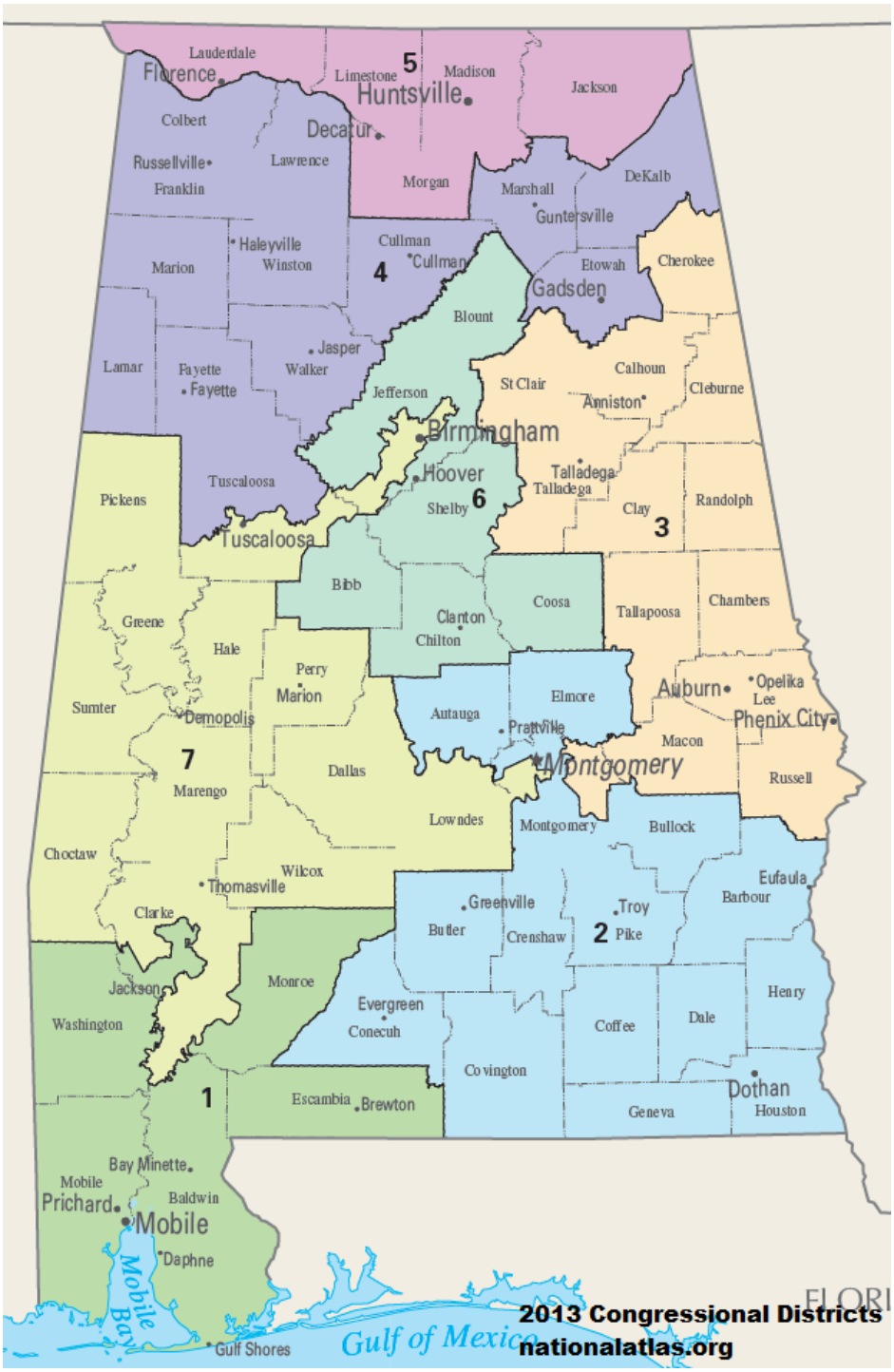
Emerson College Polling. (2017). *November 13, 2017 Alabama Senate Poll* [data file, press release, and disclosure form]. Retrieved from http://www.emerson.edu/communication-studies/emerson-college-polling-society/latest-polls.

Emerson College Polling. (2017). *Polling Society: Alumni & Student Research*. Retrieved from http://www.emerson.edu/communication-studies/emerson-college-polling-society/alumni-student-research.

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. Behavior research methods, 47(3), 685-690.

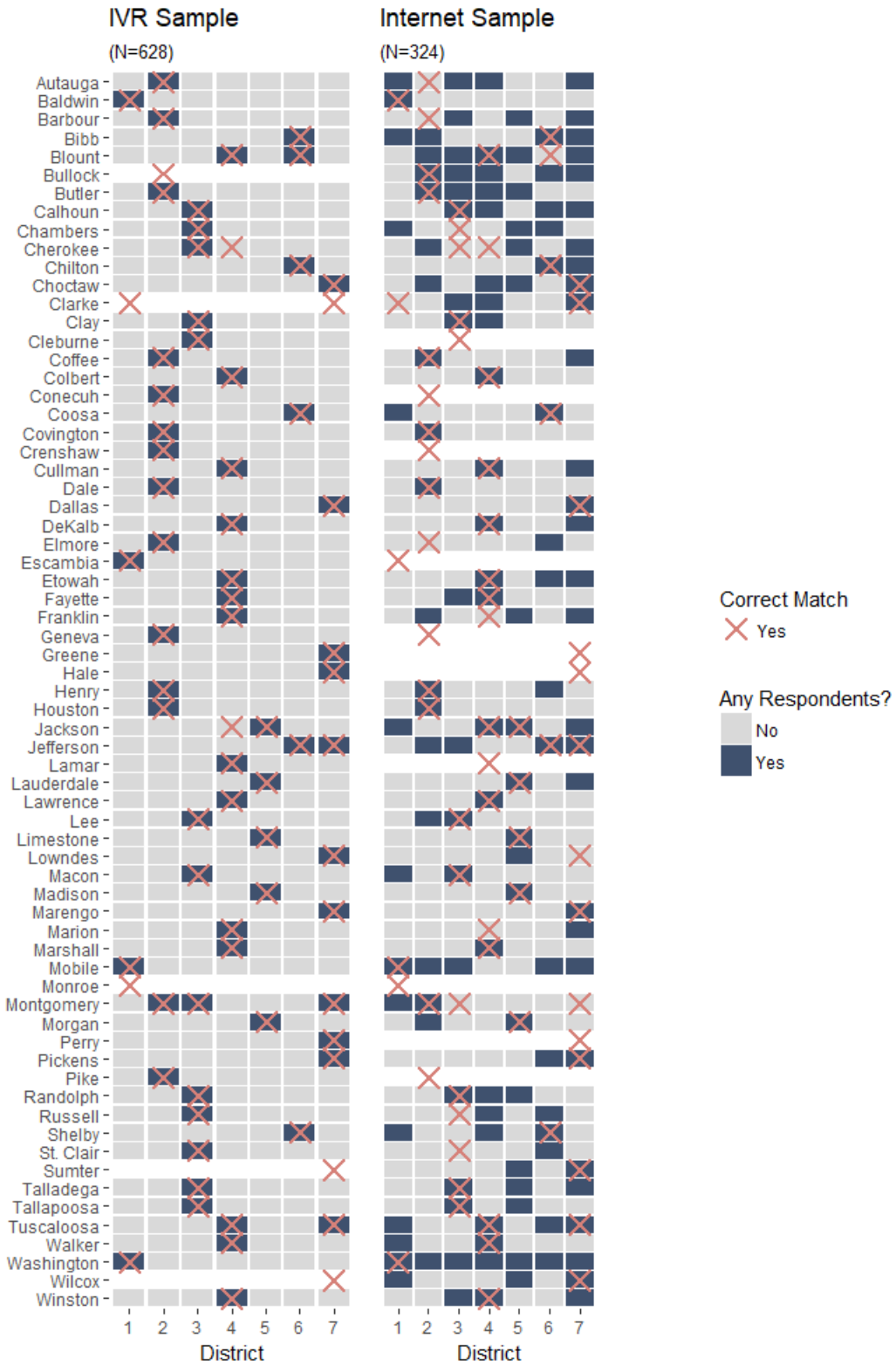
Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). The science of web surveys. Oxford University Press.

Figure 1. Alabama County and Congressional District Boundaries



Source: Department of the Interior. (2014). National Atlas of the United States. “Map of Congressional Districts in the state of Alabama, reflecting district boundaries current to the 113th United States Congress.” Retrieved on December 1, 2017 from “https://commons.wikimedia.org/wiki/File:Alabama_Congressional_Districts,_113th_Congress.tif”

Figure 2. Heat Map Depicting Joint Distribution of Counties of Residence and Congressional Districts for Respondents in the Internet and IVR Samples.



Notes: Correct Match refers to valid/logical matches for counties and congressional districts. All other cells represent invalid/illogical county-district pairs. Blue cells refer to whether one or more respondents indicated that they lived in the corresponding county and congressional district.