Improving Practices for Asking Sensitive Questions in Surveys

Pre-analysis plan

Mattias Agerberg and Marcus Tannenberg

15 januari 2020

Abstract

Social scientists regularly ask sensitive questions in surveys. This is often necessary to investigate research questions of great importance. However, respondents might be uncomfortable answering such questions directly. Researchers have therefore developed certain techniques to ask sensitive questions indirectly, in a way that is not intrusive. The so called *list experiment* is arguably the most popular among these tools. While this method is very promising, it also has several important drawbacks. In particular, the technique works poorly in the face of inattentive respondents - a problem that is greatly pronounced when vast amounts of survey data nowadays is collected online. This project will develop and test diffferent strategies, grounded in previous research, to minimize the negative effects of respondent inattention. The study we propose will allow us to give concrete recommendations to applied researchers about best practices with regard to sensitive questions in surveys.

Introduction

Many of the most interesting questions a social scientist might investigate via survey questions are *sensitive*. These are questions that respondents might be uncomfortable answering directly when the questions, for instance, touch on "taboo" topics or ask the respondent to reveal information that could be considered "too personal". This puts the researcher in a tricky situation. To be able to address important problems in society we want to know things like: who is most likely to abuse drugs? How common is bribery in a country? What do people living under a repressive regime really think of their rulers? To study topics like these, surveys of the general population are indespensible. However, it is easy to see why respondents might be reluctant to reveal such information directly.

The so called *list experiment* is a method to address this exact problem. This is a survey technique that has been used to estimate the prevalence of sensitive behavior like substance abuse, cheating, and vote buying, where the respondent does not have to directly disclose any information about the sensitive item (Glynn 2013). The list experiment works by aggregating a sensitive item (like a question about drug abuse) with a list of non-sensitive items so that the respondent only has to indicate the *number* of items that apply and not which specific items that are true. This way, the anonymity of the respondent is completely protected and his or her individual response not identifiable. However, by giving the list *without* the sensitive item to a control group the average response to the sensitive question is identifiable to the researchers. This makes it possible to measure the share of affirmative answers to a wide range of important and sensitive questions. In recent years, the use of this technique has therefore grown rapidly in the social sciences (see Blair, Coppock, and Moor (2018) for an overview).

While the list experiment is an important tool that holds great promise, the method is not without its drawbacks. A disadvantage of the standard design is that the obtained estimate of the sensitive item tends to have high variability - a consequence of the fact that the estimate is obtained by aggregating the sensitive item with a list of non-sensitive items, where the sensitive item is only given to half of the respondents (the treatment group). This has spurred a large body of work on efficient statistical estimation of the quantity of interest (Aronow et al. 2015; Blair and Imai 2012; Corstange 2009; Imai 2011; Tian et al. 2017).

However, all estimation techniques are typically sensitive to different types of respondent error (Ahlquist 2018; Blair, Chou, and Imai 2019), where some share of respondents provide inaccurate or hasty responses. While some types of error can be limited by designing the lists in a well-thought-out manner (Glynn 2013), other types of error are harder to reduce. An especially insidious type of error is so called *non-strategic* respondent error that arises when respondents provide a *random* response to the list experiment. Given that the list lengths differ between the treatment and control group (by design), this type of error can be very problematic and increase both bias and variance in the estimate of interest (Ahlquist 2018).

Non-strategic respondent error is likely to be high when respondents do not pay enough attention to the survey (Berinsky, Margolis, and Sances 2014). As many survey experiments nowadays are conducted using online platforms like MTurk, making sure respondents actually provide meaningful responses has become increasingly difficult. Some respondents might - unbeknownst to the researcher - simply rush through the questionnaire without actually engaging with the content. While the issue of low respondent effort and attention is well-known, researchers often do not consider it when analyzing their experiments (Harden, Sokhey, and Runge 2018). Given the challenges to efficient analysis of list experiments in the first place (see Blair, Coppock, and Moor (2018)), we should expect these issues to be especially pronounced here. Recent research suggests that this might indeed be the case (Alvarez et al. 2019).

This project aims at developing and testing design-based solutions and recommendations to minimize respondent error in list experiments. As argued above, minimizing this type of error is absolutely crucial for researchers to be able to study sensitive topics of importance in a manner that guarantees respondents' integrity and yields honest and useful responses. In short, the project will explore different ways to improve the average response quality among respondents. More specifically, we explore a number of techniques from previous research to raise the average attentiveness in the sample, including instructive manipulation checks (Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009), and a "warning message" to increase respondent attentiveness (Clifford and Jerit 2015). We also extend existing methods by developing a factual manipulation check for the list experiment and techniques to include a placebo item in the control list.

To be able to evaluate the different methods and criteria for excluding inattentive respondents, we develop a method where we design several list experiments where the item of interest (the "sensitive" item) has three specific properties: (1) the true quantity of the item is known, (2) the item is independent of all items on the control list, (3) the item is independent of all (observed and unobserved) respondent characteristics. We construct "sensitive" items that meet these criteria by randomly selecting an item from a list of items for each individual respondent. This way, the expected prevalence for the "sensitive" item on the list is known by design. We then compare different methods and criteria by estimating the root mean squared error of the prediction for the item of interest (a quantity we define below).

This project contributes to the literature on survey methodology in general and to the growing literature on list experiments in particular. By evaluating different strategies to improve the average response quality among respondents and suggest concrete recommendations for applied researchers, the project will benefit the research community writ large.

Methods to reduce non-strategic respondent error in list experiments

In this section we briefly describe the different methods to reduce respondent error that we consider in the project. These are methods that have shown promise in previous research but that have not been evaluated in the context of list experiments specifically. First, we consider standard "manipulation checks" (or "screeners"). A common type of manipulation check is the instructional manipulation check (IMC) (Oppenheimer, Meyvis, and Davidenko 2009). IMCs work by instructing respondents to show that they are paying attention. This is done by giving respondents a precise set of instructions to follow when responding

to the IMC items which are embedded in the survey. Respondents failing to follow the instructions are classified as "inattentive" and potentially excluded from the data analysis (Berinsky, Margolis, and Sances 2014). We also consider a second type of manipulation check that instead asks objective questions about key elements in the experiment. This type of manipulation check, referred to as a factual manipulation check (FMC), thus aims to identify individual attentiveness to experimental information directly (Kane and Barabas 2019). Below we design an FMC that is specifically tailored to the list experiment.

Another option to potentially improve the estimate of interest is to include a *placebo* item in the control list that equalizes the lists' length. If respondents answer in a manner that is correlated with the list length, for instance by using the total number of items on the list as a "reference point" (De Jonge and Nickerson 2014), this will bias standard estimators for the prevalence of the sensitive item. Equalizing the lists thus removes the bias created by some forms of non-strategic respondent error, something that we discuss further below. A placebo item in this sense is an item where the true population quantity is zero. The item should hence only increase the length of the control list, without changing the expected number of affirmative responses to the control items in the population (Riambau and Ostwald 2019). This strategy theoretically eliminates some types of bias emanating from the different list lengths. However, good placebo items can be difficult to design and implement. We discuss this problem and propose novel solutions below.

Both the IMC and the FMC can be used to identify shirking respondents. It is not, however, always obvious what to do with these. Excluding a large number of respondents might be problematic - especially if done in a careless manner (Aronow, Baron, and Pinson 2019; Berinsky, Margolis, and Sances 2014). A final strategy we consider is therefore to try to *increase* the average respondent attentiveness. A simple intervention explored by Clifford and Jerit (2015) is to provide a "warning message" to the respondents. This is a short message stating that responses are carefully checked and that only responses from participants that demonstrate that they have read and understood the survey will be used. The respondents also have to indicate that they have understood the instructions. The authors find that respondents given this message (referred to as "audit" in the study) are substantially more attentive than respondents in the control group who received no message.[1]

Illustrating the problem of measurment error in list experiments

Before turning to the empirical study, we provide some simulation evidence to demonstrate the potential effectiveness of the methods discussed above. The list experiment works by aggregating a sensitive item with a list of control items to protect respondents' integrity (Glynn 2013). We adopt the notation in Blair

[1]The authors try several different warning messages but find the "audit" message to be the most effective.

and Imai (2012) and consider a list experiment with J binary control items and one binary sensitive item denoted $Z_{i,J+1}$. Respondents are randomly assigned to either a control group ($T_i = 0$) and given the list with J control items, or to a treatment group ($T_i = 1$), given the list with with J control items *and* the sensitive item $Z_{i,J+1}$. The total number of items in the treatment group is thus $J + 1$. Y_i denotes the observed response to the list experiment for respondent i (the total number of items answered affirmatively). If we denote the total number of affirmative answers to J control items with Y_i^*, the process generating the observed response can be written as:

$$Y_i = T_i Z_{i,J+1} + Y_i^*$$

Blair and Imai (2012) show that if we assume that the responses to the sensitive item are truthful ("no liars") and that the addition of the sensitive item does not alter responses to the control items ("no design effects"), the proportion of affirmative answers to the sensitive item, denoted τ, can be estimated by taking the difference between the average response among the treatment group and the average response among the control group (i.e. a difference-in-means estimator (DiM)).[2]

Inattentive respondents potentially violate both assumptions if they provide a *random* response to the list experiment. For illustration, we simulate one basic method to decrease non-strategic respondent error: excluding respondents who fail manipulation checks. The basic simulation assumes 2000 respondents. The control list consists of 4 independent items, each drawn from a Bernoulli distribution. The parameter p_j was set to 0.5, 0.5, 0.15, and 0.85 for the different items respectively. In line with current recommendations (Glynn 2013), two of the control items were generated to be negatively correlated ($r = -0.6$). The treatment list consists of the same 4 items plus a "sensitive" treatment item, drawn from a Bernoulli distribution with $p = \frac{1}{6}$. This basic setup is very close to the setup used in Ahlquist (2018) and Blair, Chou, and Imai (2019). In the appendix we discuss and show how changes to the setup affect the simulation results.

25% of the total number of "respondents" were randomly assigned to be "inattentive" (for instance, Alvarez et al. (2019) estimates that 36% of respondents in their experiment were inattentive, and from an earlier application of the list experiments in a similar setting we estimate 28% of respondents to be inattentive (Robinson and Tannenberg 2019)). Throughout the paper we denote the share of inattentive respondents with s. We simulate a process where inattentive respondents answer randomly: for inattentive respondents in the control group the outcome was replaced by a draw from a discrete uniform distribution, $U\{0, 4\}$, and

[2]The difference-in-means estimator can be written as:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} T_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - T_i) Y_i$$

where $\hat{\tau}$ is the estimated proportion affirmative answers to the sensitive item, $N_1 = \sum_{i=1}^{N} T_i$ is the size of the treatment group and $N_0 = N - N_1$ is the size of the control group.

the outcome for the inattentive treatment group was replaced by a draw from $U\{0, 5\}$. Blair, Chou, and Imai (2019) argue that this is a plausible model for the behavior of inattentive respondents answering the list experiment. We refer to this as the *uniform error model*. If we set W_i equal to 1 to indicate that a specific respondent is inattentive (0 otherwise), the process generating the observed response under this model can be written as:

$$Y_i = (1 - W_i)(T_i Z_{i,J+1} + Y_i^*) + W_i(T_i U_{\{0,J+1\}} + (1 - T_i)U_{\{0,J\}})$$

The conseqeunces of this error process in terms of bias in the estimate of τ are similar to several other plausible error processes. We discuss this further in the appendix.

To simulate a successful manipulation check we then randomly exclude a share of the "inattentive" respondents. This share is denoted by s'. Figure 1 shows the distribution of $\hat{\tau}$ based on 10000 simulated data sets (assuming the setup described above) for different s': no inattentives excluded ($s' = 0$; purple), an ineffective attention check ($s' = 0.3$; teal), and an effective attention check ($s' = 0.8$; yellow). As shown in figure 1, the mean $\hat{\tau}$ is just above 0.266 when no inattentive respondents are excluded (purple). Considering that we are trying to estimate an item with a prevalence of $\frac{1}{6}$, this is a full 10 percentage points off from the true prevalence.[3] Excluding inattentive respondents can clearly be very beneficial under this setup. For instance, excluding 80% of the inattentive respondents (yellow) lowers the mean $\hat{\tau}$ to 0.193, less than three percentage point off from the true prevalence. Overall, this basic simulation shows that inattentive respondents in list experiments can be hugely problematic, but also that the estimates can be greatly improved by decreasing the share of inattentives in the sample.

Research design

We will explore the question of how to improve the average response quality in list experiments by fielding an original survey in the US. Given that most online studies, including list experiments, are run using samples from the US (see Berinsky, Margolis, and Sances (2014)), we consider this the most relevant population for our study. Based on several power analyses, we aim for a final sample of about 5000 respondents to be able to measure the quantities of interest with enough precision. We are collaborating with the online platform *Luc.id* for the data collection. The goal is to carry out our study during the early fall of 2020.

As described above, our study will evaluate several strategies to increase average respondent attentiveness in data from list experiments as well as ways of reducing the negative effects of respondent inattentiveness. The basic research design consists of a set of two list experiments for which the item of interest (the "sensitive"

[3] As shown by Blair, Chou, and Imai (2019), under the uniform error model the bias in the DiM estimator will be $s\,(\frac{1}{2} - \tau)$, which amounts to 0.1 in the simulation example.

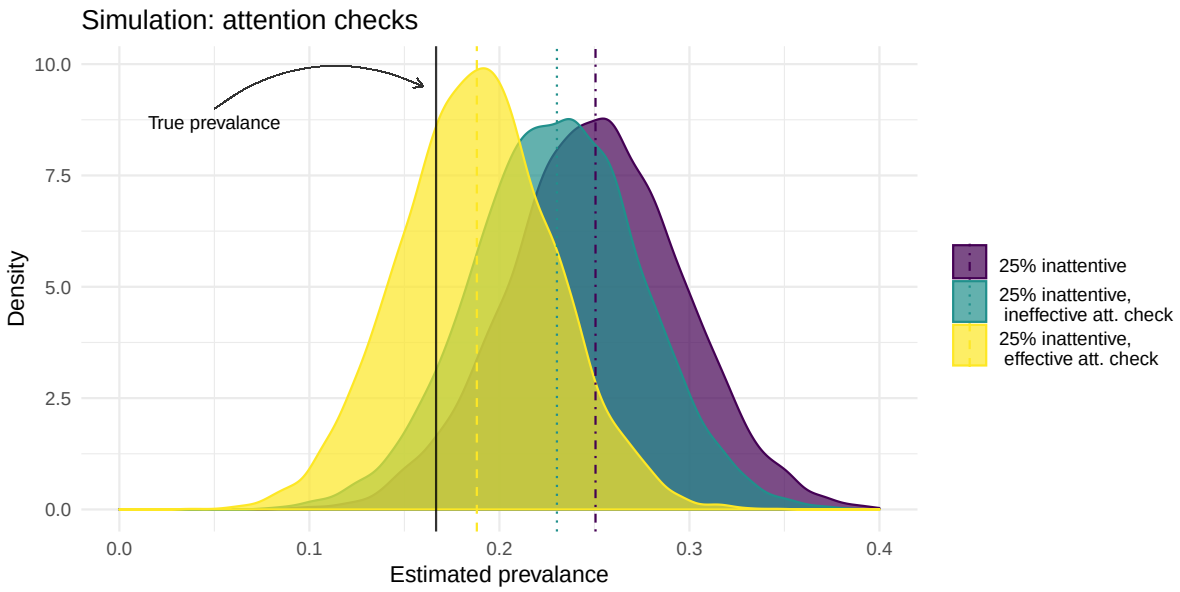

Figure 1: Simulation of $\hat{\tau}$ by exclusion of inattentive respondents. 10000 simulated data sets.

item) has three specific properties: (1) the true quantity of the item is known, (2) the item is uncorrelated with all items on the control list, (3) the item is uncorrelated to all (observed and unobserved) respondent characteristics. This way, we can calculate the bias and variance of different estimates for the item. Moreover, we know that the expected prevalence of the item does not change with the exclusion of some respondents or with the addition of a specific control item. We compare the effectiveness of different strategies to minimize respondent error by estimating the root mean squared error of the prediction (pRMSE) for the "sensitive" item:

$$pRMSE(\hat{\tau}) = \sqrt{Var(\hat{\tau}) + (\tau - \hat{\tau})^2}$$

where τ is the true quantity of interest (the true prevalence of the item in the population) and $\hat{\tau}$ is the estimated quantity from the list experiment. This statistic hence captures the trade-off between bias and variance in the list experiment. Some strategies might reduce bias but at the same time increase the variance of the estimate. Excluding inattentive respondents could, for instance, have this effect.

Both list experiments consist of a set of four control items, where the order in which they appear on the list is randomized (see full survey in appendix for details). Following best practice, two of the items on each list are negatively correlated (Glynn 2013).[4] The order in which the two lists appear are also randomized to guard against potential "learning effects" which otherwise risk limit the inferences we can draw from

[4]The combination of control items on both lists have previously been used in a similar context and does appear to work well to avoid ceiling and floor effects (see Robinson and Tannenberg (2019)).

comparing the various conditions.

For the first list experiment (A), the control list is randomly given to half of the respondents. The other half receives the treatment list, which consist of one additional, "sensitive" item. This item should have the three features described above. In the first list we use a statement about respondents' birth-timing within the year to construct a "sensitive" item with the three desired properties. The respondents who are assigned the treatment list will receive a statement of being born in one of the four seasons, for example, *I was born in December, January, or February.* The months will be randomly drawn from the four seasons and piped into the list. Agreement with the proposed statement (τ) will therefore be one quarter (25 percent) in expectation. This follows from the fact that exactly one out of the four potential statements will be true for each respondent. The true population quantity of the item will thus be known (1). Since whether or not the item is true for a given respondent is random by construction, the proposed item will also have property (2) and (3).

In the second list experiment (B), one third of the sample is randomly assigned the basic four item control list. Another third receives the control list *plus* a placebo item, for which the true prevalence by design is 0 (this is described in detail below). The remaining third receives the control list plus a treatment item. We deploy the same strategy for the treatment item as for list A. However, for list B we instead use the statement *My mother was born in January or February*, where the months will be randomized accross the six possible pairs over the year. Hence, agreement with the proposed item (τ) is two twelfths (16.66 percent) in expectation for each respondent. Again, the population quantity is therefore known (1), and given that the specific animal combination is presented at random, the item will be uncorrelated to all the control items on the list (2), as well as any respondent characteristics.

We will use the two list experiments to evaluate the effectiveness of different strategies to improve average respondent attentiveness in the sample by comparing the pRMSE between different methods. As stated above, we will implement different versions of the IMC and FMC that are specifically tailored to the setting of a list experiment. We will also test the effectiveness of a placebo item in experiment B, and evaluate the potential of actually improving respondent attentiveness with the audit message. The results will allow us to make clear recommendations to applied researchers about best practices with regard to the measurment of sensitive items in surveys. We belive this will be of great benefit to the overall research community.

References

Ahlquist, John S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators." *Political Analysis* 26: 34–53.

Alvarez, R Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis*, 1–18.

Aronow, Peter M, Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." *Political Analysis*.

Aronow, Peter M, Alexander Coppock, Forrest W Crawford, and Donald P Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology*, 43–66.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53.

Blair, Graeme, Winston Chou, and Kosuke Imai. 2019. "List Experiments with Measurement Error." *Political Analysis*.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2018. "When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments." *Working Paper*.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77.

Clifford, Scott, and Jennifer Jerit. 2015. "Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?" *Public Opinion Quarterly* 79 (3): 790–802.

Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers? Modeling the List Experiment with Listit." *Political Analysis* 17: 45–63.

De Jonge, Chad P Kiewiet, and David W Nickerson. 2014. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* 36 (3): 659–82.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77: 159–72.

Harden, Jeffrey J, Anand E Sokhey, and Katherine L Runge. 2018. "Accounting for Noncompliance in Survey Experiments." *Journal of Experimental Political Science*, 1–4.

Imai, Kosuke. 2011. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association* 106 (494): 407–16.

Kane, John V, and Jason Barabas. 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63 (1): 234–49.

Oppenheimer, Daniel M, Tom Meyvis, and Nicolas Davidenko. 2009. ""Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45: 867–72.

Riambau, Guillem, and Kai Ostwald. 2019. "Placebo Statements in List Experiments."

Robinson, Darrel, and Marcus Tannenberg. 2019. "Self-Censorship of Regime Support in Authoritarian States: Evidence from List Experiments in China." *Research & Politics* 6 (3): 2053168019856449. https://doi.org/10.1177/2053168019856449.

Tian, G.-L., M.-L. Tang, Q. Wu, and Y. Liu. 2017. "Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question." *Statistical Methods in Medical Research* 26 (2): 931–47.