Reducing Measurement Error in List Experiments

Pre-analysis plan

Mattias Agerberg and Marcus Tannenberg

22 Oktober 2019

Abstract

The *list experiment* is one of the most important tools in social science for eliciting truthful responses to sensitive questions. Under some basic assumptions the list experiment can provide an unbiased estimate of the share of affirmative answers to a sensitive question in the population of interest. A drawback of the standard design is that this estimate tends to have high variance. Recent studies also suggest that non-strategic respondent error stemming from inattentive respondents can be especially problematic in list experiments. Inattentive respondents can both further increase variance and substantially increase bias in the estimate of interest. This project aims at developing and testing design-based solutions and recommendations to minimize respondent error in list experiments. We explore four different techniques to increase the average respondent attention in the sample: Instructional manipulation checks; Factual manipulation checks; the inclusion of a Placebo Statement in the control list; and the inclusion of an Audit Warning to increase respondent attentiveness. We discuss the upsides and challenges with respect to each strategy as applied to the list experiment specifically. To empirically evaluate the different methods we design several list experiments where the true population quantity of the item of interest (the "sensitive" item) is known. We use these to test the accuracy of the list experiment in estimating the known quantity, and to explore the potential of different methods to improve the quality of provided responses. We end by discussing recommendations for applied researchers.

Introduction

In recent years, the *list experiment* has become one of the most important tools in social science to elicit truthful responses to sensitive questions. Under some basic assumptions, the list experiment can provide an unbiased estimate of the share of affirmative answers to a sensitive question in the population of interest. A drawback of the standard design is that this estimate tends to be quite variable - 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). *Strategic* respondent error in list experiments, where the respondent for instance might avoid selecting the maximum or minimum number of items, can generally be minimized by choosing control items (non-sensitive items) in a well-thought-out manner (Glynn 2013). *Non-strategic* respondent error, on the other hand, arises when respondents provide a *random* response to the list experiment. Given that the list lengths differ between the treatment and control group, this type of error will often be correlated with the treatment and can hence dramatically 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. 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) for an overview), 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. In particular, 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 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.

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 study at hand. 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]

Setup and some simulation evidence

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 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

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

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 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

[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.

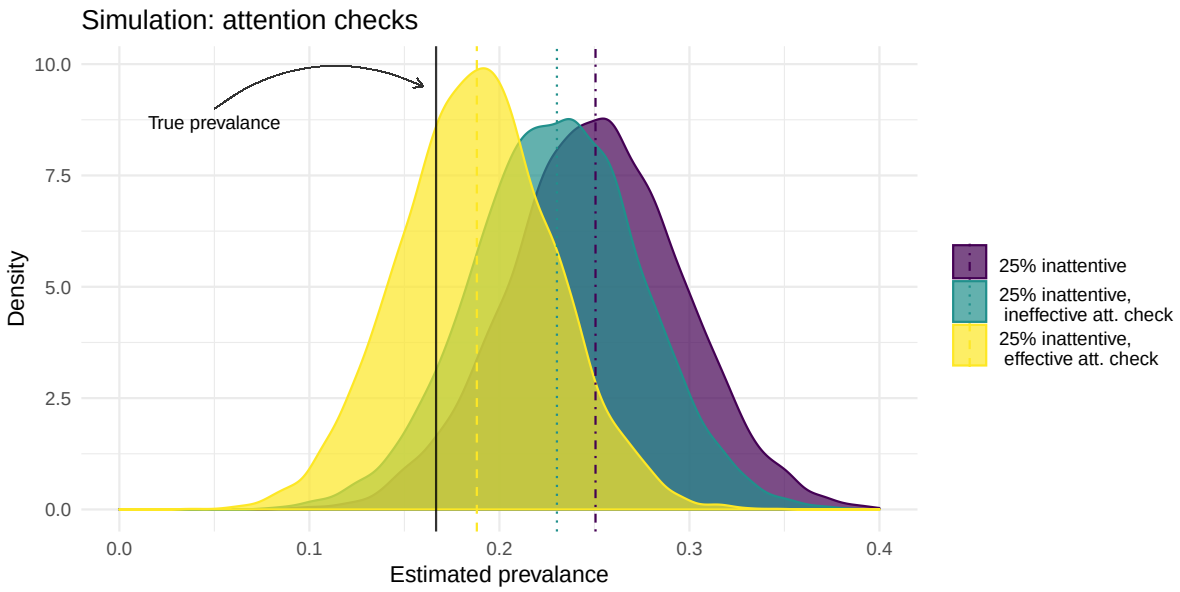

Figure 1: Simulation of $\hat{\tau}$ by exclusion of inattentive respondents. 10000 simulated data sets.

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

This study evaluates 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" 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

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

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 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 Winter (Dec/Jan/Feb).* The statement is 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. In list B we employ regarding ones zodiac animal, for example *I was born in the year of the Dog or in the year of the Pig.*, as the "sensitive" item. The study will be fielded in China, where respondents' knowledge of their zodiac animal is safe to assume. Each given year is associated with one animal of which there are 12 in total. The specific animal combination presented on the list to each respondent is randomly drawn from a list of 6 different combinations and is piped into the question item (see full survey in the appendix for the 6 zodiac statements). Hence, agreement with the proposed item (τ) is two twelfths (16.66 percent) in expectation for

[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)).

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.

The study will also include three additional list experiments, C, D, and E applied to three sensitive items: C) "I have trust in the national government in Beijing"; D) "I have trust in the local government"; and E) "I have witnessed an act of corruption or bribe-taking by a government official in the past year". List C, D and E will have same set up as list B, with one third of the sample receiving the control list, one third the control list plus a placebo, and one third the treatment list. Although these lists cannot be used to directly test the suggested solutions to minimize respondent error by comparing $pRMSE(\hat{\tau})$ (because the sensitive items do not fulfill the properties (1), (2) or (3)), the lists still allow us to test the face validity of some of the proposed approaches. For example, for the sensitive item of having *witnessed corruption* for which we expect prevalence level to be fairly low,[5] the estimate $\hat{\tau}$ obtained when the quantity is estimated using control group 1 (with the placebo item) should be *lower* than the estimate obtained when using control group 2 (no placebo item) due to mechanical inflation in the latter. Similarly, we would also expect the $\hat{\tau}$ obtained using control group 2 (no placebo item) to decrease if we exclude inattentive respondents based on the IMC and/or the FMC as long as the prevalence of the sensitive item is below 0.5. From assuming inattentive respondents to select their response at random (according to a uniform distribution) we expect the $\hat{\tau}_{inattentives} \approx 0.5$, thus their inclusion likely bias $\hat{\tau}$ upwards.[6]

Dealing with inattentive respondents

We evaluate four different strategies to improve the average respondent attentiveness in the sample: the IMC, the FMC, the inclusion of a placebo item, and a warning message. The specific questions and manipulations we use are described below.

Instructional Manipulation Check (IMC)

First we will use standard Instructional Manipulation Checks (IMC) to identify inattentive respondents. All respondents will receive two IMCs adopted from Berinsky, Margolis, and Sances (2014) and Berinsky, Margolis, and Sances (2019), one "multiple choice screener" and one "grid screener". These are minimally modified

[5] In the latest round of the Asian Barometer only 6.7 percent of respondents report that that have witnessed corruption in the last year [CITE]. While we believe this is a sensitive item that suffers from underreporting the true prevalence rate should be well below 50 percent.

[6] This follows from the fact that the expected value of a (discrete) uniform distribution is $\frac{a+b}{2}$, where the interval $[a, b]$ denotes the support of the distribution. For a list experiment with J control items and one sensitive item where respondents answer according to a uniform distribution we get $\mathbf{E}[Y_i|T_i = 1] = \frac{0+(J+1)}{2}$ for the treatment group, and $\mathbf{E}[Y_i|T_i = 0] = \frac{0+J}{2}$ for the control group. The expected difference between the groups is thus $\mathbf{E}[\tau_{inattentives}] = \mathbf{E}[Y_i|T_i = 1] - \mathbf{E}[Y_i|T_i = 0] = \frac{1}{2}$.

to suit our survey mode (see appendix for full question wording). The first IMC requires respondents to demonstrate that they are paying attention by following a precise set of instructions about which alternative or action to select at the end of a somewhat lengthy question (placed among the background questions) (Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009). In second IMC (the grid screener) the respondent is asked "In the grid below you will see a series of statements. please tell us whether you agree or disagree with each statement." The grid contain seven statements in total out of which five are sincere attitudinal statements. The remaining two probe for attentiveness and read: "Please click the"neither agree nor disagree" response" and "Two is greater than one", which both have a single correct answer. The order of the seven statements is randomized.

We treat respondents who fail either one of the IMCs as inattentive. Note that for passing the second IMC respondents have to pass both screener statements in the grid. Excluding inattentives who fail IMCs placed *before* the list experiment will not introduce post-treatment bias. It may however affect the representativeness of the overall sample, which we will explore in the paper.

Factual Manipulation Check (FMC)

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). We develop a FMC specifically tailored to the list experiment and are mindful to construct one that is not correlated with the treatment itself (receiving the list with the item of interest) so as to minimize the risk for post-treatment bias of analyses that conditions on the FMC (see Aronow, Baron, and Pinson (2019)). The FMC reads:

> You have just answered a question by selecting a number of items from a list that you agree with. Below are four items. Only one of these were featured on the previous list. Which one was it? Please select this item irregardless of if this was one of the items you agreed with or not.

To pass the FMC respondents will have to pick out the correct item from a list items that are different in character from most items on the list. We will also include an "I don't remember" response option to minimize guessing among inattentive respondents. For example, for list A the options are "We should focus less on the economy and more on the environment; I like pineapple very much; Our city needs another amusement park; I live in an apartment building;", presented in a randomized order, and with the option "I don't remember" presented last.

Our assumption is that this task is equally difficult regardless of having seen a list with 4 (the control

list) or 5 items (the treatment list). Since the FMC is administered *after* the list experiment, dropping respondents based on passage of the FMC runs the risk of introducing post-treatment bias if the character of the list (treatment or control) would affect the result of the FMC (Aronow, Baron, and Pinson 2019). While there is no perfect way to guard against the possibility of post-treatment bias in this case, we propose two tests to check if the assumption of *no* post-treatment bias is plausible. First, we regress a binary variable y that equals 1 for respondents passing the FMC (0 otherwise) on an indicator variable T that equals 1 for respondents assigned to the treatment list and 0 otherwise. We use logistic regression to estimate the equation and compute a LR-test to see if the inclusion of T is an improvement over the null-model. A rejection of the null-hypothesis in the LR-test would be evidence that the passage rates on the FMC are not equal for the treatment and control group in the list experiment. Second, we estimate the following logistic regression model *after* excluding respondents based on failure on the FMC: $logit(T_{FMC=1}) = \mathbf{X'}_{FMC=1}\beta + \epsilon$, where $T_{FMC=1}$ is the treatment assignment in the list experiment, $\mathbf{X}_{FMC=1}$ is a vector of pre-treatment covariates, and ϵ is the error term. We again compute a LR-test to test the null-hypothesis that the coefficients in the vector β are jointly equal to 0. This hence amounts to a "balance test": does the exclusion of respondents due to failing the FMC create imbalances on observed covariates between the treatment and control group in the list experiment? As noted by Aronow, Baron, and Pinson (2019), tests like these can merely give us an indication by (ideally) not providing positive evidence in favor of post-treatment bias. However, our experimental design also gives us the possibility to simply check what happens when we condition on the FMCs: does this decrease bias in the estimate of τ? Together, these tests and explorations allow us to evaluate the effectiveness of the FMCs and whether these are likely to introduce post-treatment bias.

Reducing list effect bias - placebo statement

Another option to potentially reduce the bias of the estimate of interest is to include a *placebo* item in the control list that equalizes the lists' length. Equalizing the lists' length is a way to guard against potential *list effect bias* stemming purely from the fact that one list (the treatment list) is longer than the other. Even (somewhat) attentive respondents may interact with the list differently only depending on whether they are assigned to a list with 4 or 5 items. In the face of list effect bias Y_i^* might both be the function of the true answers to the individual control items *and* a function of the list's length L_J. For instance, some respondents may use the number of items on the list to figure out how many items on the list that "should" apply to them, for example by anchoring at the mean number of items of the list, and then adjust from this value. Or some respondents might use the maximum response to determine if a certain low response (say 1) is "too low". In both these cases a respondent might adjust a true low response upwards. In this sense the total

number of items on the list could work as a "reference point" for respondents (see De Jonge and Nickerson (2014) for on overview of the psychological literature on such effects). We refer to such an effect, where the respondent's reported answer to the list deviates from their true aggregated preference for the items on the list, as a *list effect*.

In general, such list effects are not a problem as long as the effect is the same for the treatment and control group ($L_J = L_{J+1}$). Given that many of the proposed effects are directly related to the number of items on the list (for instance, anchoring using the mean number), it is not unreasonable to assume that $L_J \neq L_{J+1}$. This would create list effect bias in the estimate of τ equal to ΔL, where $\Delta L = L_{J+1} - L_J$.[7] In the examples above ΔL would be positive, leading to an artificially inflated estimate of τ. Introducing a placebo item on the control list thus makes $\Delta L = L_{J+1} - L_{J+1} = 0$, assuming that L_{J+1} is independent of both the placebo item and the sensitive item. To obtain an unbiased estimate of τ with the DiM estimator the expected prevalence of the placebo item for each respondent has to be 0.

Important to note is that list efect bias is different from the bias created by inattentive respondents. In general, a placebo item is not necessarily an optimal way of reducing bias stemming from inattentive respondents. In the appendix we show that under the uniform error model, the inclusion of a placebo item changes the bias in the DiM estimator from $s\,(0.5 - \tau)$ (no placebo) to $-s\tau$ (placebo). Rather, a placebo item has the potential to equalize any list effects between the treatment and control group among *attentive* respondents.

Previous studies considering techniques to equalize the list length between the treatment and the control group have not sufficiently discussed the different implications this might have for inattentive and attentive respondents (De Jonge and Nickerson 2014; Holbrook and Krosnick 2010; Riambau and Ostwald 2019; Tsuchiya, Hirai, and Ono 2007). Consider two different control lists, C_J and C_{J+pl}. C includes J control items and C_{J+pl} includes the same set of control items plus a placebo item (where the true prevalence for each respondent is 0). For respondents answer according to the uniform error model the expected difference between these lists is $s/2$. This follows from the fact that the expected value of τ for inattentive respondents under this error model is 1/2 (see above). Any test of potential list effect bias should therefore first try to exclude all truly inattentive respondents; otherwise we have no way of distinguishing list effect bias from bias created by inattentive respondents. Testing for list effect bias would thus involve estimating $\mathbf{E}[Y_i|C_{J+pl}] - \mathbf{E}[Y_i|C_J]$, using the sample of attentive respondents. If the estimated quantity is different from 0, this is evidence in favor of list effect bias. In the presence of such bias the inclusion of a placebo item in the control list can be expected to decrease bias in the estimate of τ.

To explore whether the inclusion of a placebo item is warrented we need an item for which the true

[7] This would hence be a violation of the "no design effects" assumption.

expected prevalence for each respondent is 0. Ideally such an item should be *plausible* for all respondents, yet by design *necessarily false* for any one given respondent. We propose a design for a placebo item that can be implemented in any programmed survey, such as web-administrated or tablet-administrated surveys, where it is possible to pipe in an item utilizing information gained earlier in the survey. This can be done in a number of different ways. In our application we will give survey respondents who indicate that they are below 30 years of age the statement "I was born in the 70s", and respondents who indicate that they are 30 or above get the placebo statement "I was born in the 2000s". To the best of our knowledge this approach to assigning a placebo item is a novel innovation. Theoretically this approach has one clear benefit vis-a-vis existing approaches. For example, in the Singaporean setting Riambau and Ostwald (2019) use "I have been invited to have dinner with PM Lee at Sri Temasek [the Prime Minister of Singapore's residence] next week.", which they suggest is "plausible but false" for all respondents. The authors caution against using ridiculous items (such as having the ability to teleport oneself) so as not to risk compromise the perceived seriousness of the survey. We take this one step further and suggest that there is a benefit to having an placebo item that is *truly* plausible to signal seriousness. Using an item that is necessary true or necessary false due to *implausibility* risks signaling to the respondent that their responses are not important or valuable to the researchers, which risk result in lower attentiveness.

We will explore the inclusion of a placebo item by randomly assigning the placebo item to one third of the sample of list B, control group 1 (C_{J+pl}). Another third of the sample, control group 2 (C_J), get the basic control list containing the four items. The remaining third of the sample get the treatment list. We will use the same setup, with two control groups, for list experiment E. List experiment B is designed according to the same principles as list experiment A, with a "sensitive" item that has properties (1), (2), and (3), while list experiment E has a sensitive item for the treatment list. We will use list B and E to test for list effect bias. We will first use the IMC to exclude exclude inattentive respondents. We will then use the DiM estimator to estimate $\mathbf{E}[Y_{i,IMC=1}|C_{J+pl}] - \mathbf{E}[Y_{i,IMC=1}|C_J] = \hat{\Delta L}$. We consider an estimated quantity that is statistically different from 0 to be evidence in favor of list effect bias. We will repeat this procedure but instead using the FMCs to exclude respondents.[8] If $\hat{\Delta L}$ is estimated to be statistically different from 0 in one or more of these tests this is an indication that researchers should consider including a placebo item in the control list. If we find evidence in favor of list effect bias we will also compare the $pRMSE(\hat{\tau})$ for list experiment A when we use the two different control lists respectively. In the presence of such bias we expect this quantity to be lower when we use the placebo list. See the appendix for a discussion of statistical power of tests pertaining to the placebo item.

[8] In addition, the placebo lists can also be used as a sanity check for the IMC and the FMC: if these exclusion criteria are good at identifying inattentive respondents they should reduce the number of respondents who answer "5" on the placebo lists to near 0.

Improving attentiveness of respondents - audit check

Finally, we also test the possibility of *improving* the attentiveness of respondents in the sample. In order to do this half of the sample is randomly assigned to receive a "warning message" just before being presented with the first list experiment. The other half of the sample receive no message.[9] The message is taken from Clifford and Jerit (2015) who develop and test several types of warning messages and find that the "audit" message is the most effective in increasing respondent attentiveness. 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 full question wording reads:

> We check responses carefully in order to make sure that people have read the instructions for the task and responded carefully. We will only accept participants who clearly demonstrate that they have read and understood the survey. Again, there will be some very simple questions in what follows that test whether you are reading the instructions. If you get these wrong, we may not be able to use your data. Do you understand? [Yes, I understand; No, I do not understand]

We analyze the effectiveness of the warning message in several ways. First, we will estimate the effectiveness of the message by measuring what fraction of respondents in each group that passes the FCMs. Since the FMCs are directly related to the lists, this will give an overall indication of whether the message plausibly increases the attention respondents pay to the list experiments. We will also compare the $pRMSE(\hat{\tau})$ in the two list experiments (A and B) between the treatment group (that received the message) and the control group. Given that we are splitting the sample in half when comparing the two groups (since we analyze the list experiment separately for the two groups), this comparison has less statistical power. We discuss this more in depth in the appendix.

Expectations and main analysis

All our proposed methods aim to minimize non-strategic respondent error or the consequences of such error. As discussed above, this can be achieved by either excluding inattentive respondents (IMC and FMC), by increasing respondent attentiveness (audit), or by minimizing the bias introduced by such error (placebo item). Our guiding premise and prior are that all proposed methods *in theory* should reduce the error of the prediction. As regards how much, or which one that is most effective, we take an exploratory approach in this study.

[9] For all other analyses we simply average over this manipulation.

We will start by evaluating the effectiveness of each method individually and make basic comparisons between the methods. For the basic analysis of the list experiments we rely on the standard DiM estimator (corresponding to the linear estimator in Imai (2011)). We will implement the estimator using a linear regression model and robust standard errors (HC2). The estimator gives us $\hat{\tau}$ and $Var(\hat{\tau})$. Given that we know τ based on the construction of the experiment, these quantities can then be used to calculate $pRMSE(\hat{\tau})$. For each method we will also display the estimate for the "sensitive item" graphically, including 95% confidence intervals and a line indicating the true value of interest (τ).

We aim for a sample of 4000 to 5000 respondents. This will allow us to estimate $\hat{\tau}$ in the overall sample with good precision (see Blair, Coppock, and Moor (2018)). We discuss questions related to statistical power further in the appendix. For list experiment A we will make the following comparisons. First, we will compare the original estimate (with all respondents), the IMC, and the FMC,[10] using the whole sample. The point of this comparison is to see if we are able to reduce the $pRMSE(\hat{\tau})$ when we exclude respondents based on the IMC and the FMC, and also to see which manipulation check that does this most efficiently. In addition, we will explore the possibility of *combining* the IMC and the FMC. In theory, this strategy can potentially reduce the $pRMSE(\hat{\tau})$ the most. We will also use experiment A to give us an initial assessment of the effectiveness of the audit treatment. This will be done by estimating $\hat{\tau}$ for the half of the sample that received the warning message (around 2000 respondents) and compare this to the estimate for the sample that did not receive the message. This comparison might be somewhat underpowered, depending on how effective the audit message can be expected to be (see appendix). We therefore do not consider this specific comparison our main tool for evaluating the audit message.

In addition, we will evaluate the audit message by comparing the share of respondents who pass the FMCs between the treatment and control group. This hence constitutes a more "indirect" test of the effectiveness of the audit message. The test involves regressing an indicator variable that equals 1 for respondents passing a FMC (0 otherwise) on a variable indicating treatment status in the audit experiment (using a logistic regression model). We will analyze all three FMCs this way. The FMC for the final list experiment allows us to see if the potential effect of the audit treatment persists over the course of the survey. This test is in general well-powered: simulation evidence suggests that we have about 80% power (assuming 4000 respondents) to detect a difference of 4 percentage points between the treatment and control group for passage rates of the FMC.

List experiment B and E will allow us to evaluate the inclusion of a placebo item. As described above, these list experiments involve three groups (with around 1300 respondents each): a "normal" control group,

[10] We will analyze each FMC in relation to respondent treatment status in each list experiment to look for signs of post-treatment bias, using the procedure described above.

a control group receiving the placebo item, and the treatment group receiving the "sensitive" item. To evaluate the presence of list effect bias we will compare the two control groups as described above, after excluding inattentive respondents. If we find evidence of such bias we will also evaluate the placebo method by comparing the estimate of $\hat{\tau}$ using two different (but overlapping) samples: the normal control group and the treatment group, and the placebo control group and the treatment group. In the presence of list effect bias we expect the $pRMSE(\hat{\tau})$ to be lower when using the placebo control group.

Finally, our design allows us to test for any *learning effects* over the first two list experiments (A and B). Related to this, Rosenfeld, Imai, and Shapiro (2016) argue that *the randomized response method* benefits from allowing respondents to *practice* once before the real round. While the randomized response method might be more demanding on the respondent than the list experiment, Tsai (2010) caution the use of it in rural China and Kramon and Weghorst (2019) show the method to fault in Kenya among less educated respondents. To test for learning effects we will compare the $pRMSE(\hat{\tau})$ when the quantity is estimated using respondents who saw list A first as their first list, to when it is estimated using respondents received list A as their second list. We do the same comparison with regards to list B, while excluding the placebo control group. Again, the order in which respondents receive the two list is randomized. Potentially, the $pRMSE(\hat{\tau})$ might be lower when respondents get a specific list experiment as the *second* experiment, having "practiced" the method once.

Additional analysis intended for a separate research note

Reevaluating hierarchical trust in China and examining sub-group differences to self-censor

In addition to functioning as additional checks for the "Reducing measurement error of list experiments" paper, list experiment C, D and E are designed to help us address two issues of substantive importance within the study of political support in authoritarian regimes: namely the existence of hierarchical trust in China as well as sub-groups' sensitivity bias to political survey items. We use *sensitivity bias* and *self-censorship* interchangeably.

The first issue pertains to the longstanding observation that Chinese respondents exhibit more trust in the national government than in the local government[11] (see Li and O'Brien (1996); Bernstein and Lü (2003)). This is opposite to what we observe in most western and all other Asian countries for which there is survey data (Wu and Wilkes 2018). This anomalous pattern has been subject to much scholarly attention and, for example, been explained by the differential impact on trust in the local and national government stemming from: implementation of education reform (Lü 2014); land requisitions (Cui et al. 2015); state media consumption (Li 2004); and Confucian values (Wong, Wan, and Hsiao 2011). An alternative explanation is that this anomaly is driven by over-reporting of trust in the central government. Wu and Wilkes (2018) find that holding hierarchical trust is associated with reported political fear.[12]. The practice of state run media to scapegoat local governments and officials to improve perceptions of the government in Beijing may not only shape perceptions (Li 2004), but can also signal that distrust of your local government is a sanctioned opinion to hold. Expressing distrust of the central government, however, may not be sanctioned. Indicative of the latter, Li (2016) provides indirect evidence that trust in the central government is weaker than suggested by surveys, and in a previous study we show direct evidence of substantive self-censorship (underreporting) with regards to "Confidence in the national government" (Robinson and Tannenberg 2019). It is possible that this highly scrutinized pattern of hierarchical trust does, in fact, not exist. In this study we compare the prevalence of trust in the national government and the local government obtained through indirect estimates (from list experiment C and D). To get at potential differences in perceived sensitivity of the two items we compare the list estimates to the estimated prevalence of trust in the two levels when asked directly.

The second purpose of this study is to investigate if there are differences in the propensity to self-censor

[11] In the most recent round of the Asian Barometer just shy of 87 percent of respondents reported "A great deal" or "Quite a lot" of trust in the national government, while only 64 percent did so for the Local government.

[12] The authors use responses to "people are free to speak what they think without fear", and "people can join any organization they like without fear" to measure political fear. It should be noted that these too are sensitive items which may suffer from underreporting.

among a number of sub-groups. First, we have expectations with regards to two sets of sub-groups: gender and age. In previous work we have documented that women are more effected by sensitivity bias than are men: women report higher regime support, and lower experienced corruption when asked directly, but when asked indirectly women instead report *lower* levels of regime support and *higher* levels of experienced corruption than men (see Robinson and Tannenberg (2019) and Agerberg (2019)). The systematic variation in self-censorship is large enough to fully reverse the observed relationship between gender and the sensitive items as estimated through direct questioning. We observe the same pattern with regards to younger (below 30) and older (30 or above) respondents: with younger respondents reporting higher levels of regime support when asked directly than do older respondents, while exhibiting lower levels of support when estimated with list experiments. In addition we expect respondents with urban hukou (household registration) to be more effected by sensitivity bias than rural respondents due to the design of the Chinese social system with urban hukou holder having access to more services. Again we observe this pattern in our previous study. Second, we will conduct an explorative analysis of subgroup differences based on binary measures of income (High/Low); and education (University/No university); party membership (Yes/No). While our previous studies have been relatively underpowered to establish differences of misreporting among subgroups (as have most previous applications of list experiments (Blair, Coppock, and Moor 2018)), we are aiming for a sample size of 4000 to 5000 respondents in this study which will allow us to detect the existence of any substantial differences.

Analysis

We will use the IMC described above to exclude inattentive respondents from all analyses. In order to evaluate the extent of self-censorship we ask for agreement with the three sensitive items directly. The direct questions will be given to the full sample in order to get as precise estimates as possible. They will be placed *after* the last list experiment in order to avoid priming effects on the list experiment. This survey flow does however carry the risk that asking a treated individual the direct question post-treatment could produce priming effects. This can be tested; if presentation of the treatment list primes one to respond in a certain manner to the direct question, or to opt for a "Do not know" response, treatment status should correlate with outcome of the direct question. To test this we regress a y_{agree} that equals 1 for respondents who "agree" with the sensitive item in direct questioning and 0 otherwise on an indicator variable T that equals 1 for respondents assigned to the treatment list and 0 otherwise. We use logistic regression to estimate the equation and compute a LR-test to see if the inclusion of T is an improvement over the null-model. A rejection of the null-hypothesis in the LR-test would be evidence that the direct estimates of the sensitive item are not equal for the treatment and control group in the list experiment. We do the same for for y_{DK} for which 1 equals

"Do not know" and 0 any response to the direct item. In the event of priming effects we will simply exclude respondents who received the corresponding treatment list when estimating $\hat{\tau}_{direct}$. If we detect no priming we retain the full sample for the direct estimates.

First, we will test the extent of self-censorship for the three sensitive items: (C) "I have trust in the national government in Beijing"; (D) "I have trust in the local government"; and (E) "I have witnessed an act of corruption or bribe-taking by a government official in the past year". We compare this estimate with that obtained from the direct questioning method ($\hat{\tau}_{direct}$). We code respondents answering "do not know" as 0 (*no trust* for C, D, and *not having witnessed corruption* for E). We use the procedure described in Blair and Imai (2012) and compare the predicted responses to the direct questions, modeled with a logistic regression model, to the predicted responses to the sensitive item in the different list experiments respectively. The procedure gives us an estimate of the amount of sensitivity bias (the difference between the direct and indirect question), including 95% confidence intervals around the estimates (obtained via Monte Carlo simulations; see Blair and Imai (2012)). We consider a sensitivity bias estimate that is statistically different from 0 as evidence of self-censorship.

Second, we use the estimates of self-censorship to compare the degree of misreporting bias between the items of trust in the national and the local government. We expect misreporting bias to be larger at the national level. To compare overall levels of misreporting bias we first use the procedure described above to obtain estimates for the direct and indirect questions for national and local government respectively. This gives us an estimate of sensitivity bias for each level of government, allowing us to calculate the difference between the two estimates. We use a bootstrap procedure to compute a one-sided confidence interval for the estimated difference: we resample indices in the data set with replacement and repeat the procedure above a large number of times (> 1000). We use the estimated sampling variance of the difference to calculate a one-sided 95% non-parametric bootstrap confidence interval for the estimate. We reject the null hypothesis of no difference if the calculated interval excludes 0. Based on simulations we estimate that the true difference in reporting bias has to be around 8 percentage points for the procedure to have 80% power to reject the null hypothesis (given 4000 respondents).

Third, we will test for differences in sensitivity bias among subgroups by calculating multiple regression estimates of the list items with included covariates, using the *List* package in R (see Blair and Imai 2012). For list experiment C and D (which do not include a placebo item in the control group) we use the Nonlinear Least Squares (NLS) estimator to estimate agreement with the sensitive items when asked indirectly ($\hat{\tau}_{list}$) (see Imai (2011)), conditional on respondent characteristics. By including the same covariates when modeling the direct question we can estimate the predicted amount of self-censorship (or "social desirability bias") as a function of different respondent characteristics. We do the same for list experiment E but instead use the

standard linear estimator (Imai 2011). The reason for using the linear estimator instead of the NLS estimator is the inclusion of a placebo item in one of the control groups for list E. This procedure lets us explore which respondent characteristics that are relevant in determining self-censorship. Note that for reasons of statistical power we will only explore this *within* each specific list experiment and not make subgroup comparisons across the different experiments.

Additional analysis intended for a separate research note

Gender and under-reporting of corruption experiences

In addition to the analyses described above we will use list experiment E to test the hypothesis that women are more likely to under report personal experiences with corruption than men. This is a replication of the results from another study that we ran in Romania (Agerberg 2019). The results from that study suggest that people are likely to under report their direct experiences with corruption, on average. Moreover, the result suggest that the under reporting is much more pronounced among women - probably because women consider such questions to be more sensitive than men. This runs contrary to studies arguing that women are less likely to engage in corruption because they are less likely to get asked to pay bribes in the first place (Goetz 2007; Heath, Richards, and Graaf 2016; Mocan 2004).

Analysis

The Romanian study ($N = 3027$) used a list experiment to estimate agreement with the sensitive item "Being asked to pay a bribe to a public official" (in the past 12 months). The study also included a direct question of the same item, asked to the control group in the list experiment. When agreement with the item was estimated using the list experiment the proportion of affirmative answers was estimated to be 0.35 $[0.26, 0.44]$ (including 95% confidence intervals). When instead using the direct question, the same proportion was estimated to be 0.19 $[0.17, 0.21]$. Using the procedure described in Blair and Imai (2012), the amount of "sensitivity bias" was estimated to be 0.16 $[0.07, 0.25]$. To estimate whether the amount of sensitivity bias differs between male and female respondents we used the following bootstrap procedure: we first resample indicies in the dataset with replacement. We then use the bootstrap sample to estimate the agreement with the sensitive item from the list experiment for men and women separately, using the linear estimator in Imai (2011). We do the same for the direct question, using a logistic regression model. We then estimate $(\hat{\tau}_{list,female} - \hat{\tau}_{direct,female}) - (\hat{\tau}_{list,male} - \hat{\tau}_{direct,male})$ to calculate the difference in sensitivity bias between women and men. We repeat this procedure 10000 times to get the sampling distribution of the estimate and calculate a 90% nonparametric BCa confidence interval around the estimate (this is equivalent of a one-sided 95% confidence interval, given that we want to test the hypothesis that the sensitivity bias is larger for women). The results show that the under reporting is much more pronounced for women. The estimated difference, using the procedure described above, is 0.26 $[0.10, 0.42]$.

We will use list experiment E to replicate this result. As described above, the sensitive item in the experiment is "I was asked to pay a bribe to a government official during the past year". We will use the

IMC to exclude inattentive respondents from the analysis. The list experiment contains two control groups (placebo and normal). We will pool these into one control group that we then compare to the treatment group in the list experiment. The direct question is asked to all respondents after the list experiment. We will use the procedure described in the previous section on hierarchical trust to determine if we should use all respondents or only respondents in the control group to get the direct estimate. We will then use the exact same bootstrap procedure that we describe above with regard to the Romanian experiment to determine if under reporting is higher among women. If the 90% nonparametric confidence interval around the estimate is above 0 we consider the estimated difference to be statistically significant.

References

Agerberg, Mattias. 2019. "Corrupted Estimates? Response Bias in Citizen Surveys on Corruption." *Working Paper*.

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

Alvarez, R Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis*. Cambridge University Press, 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*. Oxford University Press, 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). Wiley Online Library: 739–53.

———. 2019. "Using Screeners to Measure Respondent Attention on Self-Administered Surveys: Which Items and How Many?" *Working Paper*.

Bernstein, Thomas P, and Xiaobo Lü. 2003. *Taxation Without Representation in Contemporary Rural China*. Cambridge University Press.

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). Oxford University Press: 790–802.

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

Cui, Ernan, Ran Tao, Travis J Warner, and Dali L Yang. 2015. "How Do Land Takings Affect Political Trust in Rural c Hina?" *Political Studies* 63. Wiley Online Library: 91–109.

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). Springer: 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.

Goetz, Anne Marie. 2007. "Political Cleaners: Women as the New Anti-Corruption Force?" *Development and Change* 38 (1): 87–105.

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

Heath, Anthony F, Lindsay Richards, and Nan Dirk de Graaf. 2016. "Explaining Corruption in the Developed World: The Potential of Sociological Approaches." *Annual Review of Sociology* 42: 51–79.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports." *Public Opinion Quarterly* 74 (1). Oxford University Press: 37–67.

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

Kramon, Eric, and Keith Weghorst. 2019. "(Mis) Measuring Sensitive Attitudes with the List Experiment: Solutions to List Experiment Breakdown in Kenya." *Public Opinion Quarterly* 83 (S1). Oxford University Press UK: 236–63.

Li, Lianjiang. 2004. "Political Trust in Rural China." *Modern China* 30 (2). Sage Publications: 228–58.

———. 2016. "Reassessing Trust in the Central Government: Evidence from Five National Surveys." *The China Quarterly* 225. Cambridge University Press: 100–121.

Li, Lianjiang, and Kevin J O'Brien. 1996. "Villagers and Popular Resistance in Contemporary China." *Modern China* 22 (1). Sage Publications Sage CA: Thousand Oaks, CA: 28–61.

Lü, Xiaobo. 2014. "Social Policy and Regime Legitimacy: The Effects of Education Reform in China." *American Political Science Review* 108 (2). Cambridge University Press: 423–37.

Mocan, Naci. 2004. "What Determines Corruption? International Evidence from Micro Data." *NBER Working Paper Series*, 10460.

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.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60 (3). Wiley Online Library: 783–802.

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.

Tsai, Lily L. 2010. "Quantitative Research and Issues of Political Sensitivity in Rural China." In *Contemporary Chinese Politics: New Sources, Methods, and Field Strategies*, edited by Allen Carlson, Mary E Gallagher, Kenneth Lieberthal, and Melanie Manion, 246–65. Cambridge University Press New York.

Tsuchiya, Takahiro, Yoko Hirai, and Shigeru Ono. 2007. "A Study of the Properties of the Item Count Technique." *Public Opinion Quarterly* 71 (2). Oxford University Press: 253–72.

Wong, Timothy Ka-ying, Po-san Wan, and Hsin-Huang Michael Hsiao. 2011. "The Bases of Political Trust in Six Asian Societies: Institutional and Cultural Explanations Compared." *International Political Science Review* 32 (3). Sage Publications Sage UK: London, England: 263–81.

Wu, Cary, and Rima Wilkes. 2018. "Local–National Political Trust Patterns: Why China Is an Exception." *International Political Science Review* 39 (4). SAGE Publications Sage UK: London, England: 436–54.

Appendix

Bias under different error models

In this section we discuss different models of non-strategic respondent error as well as the consequences of adding a placebo item to the control list under different error models. Throughout the text we generally assume that inattentive respondents choose according to what we refer to as the *uniform error model*. This error process can formally be described as follows: Let s be the share of inattentive respondents in the sample and let τ be the true proportion of respondents for whom the "sensitive" item $Z_{i,J+1}$ is true. Let Y_i^* represent the total number of affirmative answers to J control items. Respondents in the treatment group ($T_i = 1$) are given the list with J control items plus the sensitive item and respondents in the control group ($T_i = 0$) are given the list with J control items. Assume that inattentive respondents answer according to a random draw from a discrete uniform distribution ($U\{0, J\}$ for respondents assigned to the control list and $U\{0, J+1\}$ for the treatment list). W_i equals 1 if a particular respondent is inattentive (0 otherwise). Under this model the process generating the observed response can be written as:

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

As shown by Blair, Chou, and Imai (2019 appendix A), under the model the bias of the difference-in-means estimator, $\mathbf{E}[Y_i|T_i = 1] - \mathbf{E}[Y_i|T_i = 0] - \tau$, will be:

$$\left\{\mathbf{E}\left[(1-s)(Y_i^* + Z_{i,J+1}) + s\frac{J+1}{2}\right] - \mathbf{E}\left[(1-s)Y_i^* + s\frac{J}{2}\right]\right\} - \tau = (1-s)\tau + \frac{s}{2} - \tau = s\left(\frac{1}{2} - \tau\right)$$

When adding a placebo item (with expected value equal to 0) to the control list under this error model the bias instead becomes:

$$\left\{\mathbf{E}\left[(1-s)(Y_i^* + Z_{i,J+1}) + s\frac{J+1}{2}\right] - \mathbf{E}\left[(1-s)Y_i^* + s\frac{J+1}{2}\right]\right\} - \tau = ((1-s)\tau + 0) - \tau = -s\tau$$

Hence, without a placebo item the bias will be 0 only when $\tau = \frac{1}{2}$, and with a placebo item the bias will be 0 only when $\tau = 0$. The addition of the placebo item will thus in general cause negative bias ($-s\tau$) under the uniform-error model. This is because the estimated prevalence of the sensitive item in the inattentive group is 0 ($\mathbf{E}[(J+1)/2] - \mathbf{E}[(J+1)/2]$).

When does the addition of a placebo item decrease the amount of bias under the uniform-error model?

This happens when $s\left(\frac{1}{2}-\tau\right) > s\tau$. Hence:

$$s\left(\frac{1}{2}-\tau\right) > s\tau \Rightarrow \frac{1}{2}-\tau > \tau \Rightarrow \frac{1}{2} > 2\tau \Rightarrow \tau < \frac{1}{4}$$

The addition of a placebo item will hence decrease the amount of bias as long as the true prevalence of $Z_{i,J+1}$ is below $\frac{1}{4}$.

A different reasonable model of respondent error among inattentive respondents is a *binomial* model with $p = 0.5$. Under this model respondents hence answer affirmatively to each item with a 50/50 chance. Inattentives in the control group respond according to $Y_i \sim Binom(J, 0.5)$, while inattentives in the treatment group select according to $Y_i \sim Binom(J+1, 0.5)$. The bias of the DiM estimator under this error model is the same as for the uniform model:

$$\left\{\mathbf{E}\left[(1-s)(Y_i^* + Z_{i,J+1}) + s\binom{J+1}{y}0.5^y 0.5^{(J+1)-y}\right] - \mathbf{E}\left[(1-s)Y_i^* + s\binom{J}{y}0.5^y 0.5^{J-y}\right]\right\} - \tau =$$
$$(1-s)\tau + s(J+1)0.5 - sJ0.5 - \tau = (1-s)\tau + \frac{s}{2} - \tau = s\left(\frac{1}{2}-\tau\right)$$

The bias when adding a placebo item to the control list would also be the same as in the uniform case ($-s\tau$) since both the treatment and control group would respond according to $Y_i \sim Binom(J+1, 0.5)$.

A third model might be that inattentive respondents randomly select the *middle* response alternative. When J is even[13], respondents in the control group responds affirmatively to $\frac{J}{2}$ items, while respondents in the treatment group randomize between $\frac{J}{2}$ and $\frac{J}{2}+1$ items (since it is not possible to select $\frac{J+1}{2}$ items when J is even). The bias of the DiM estimator under this model is, again, the same as for the uniform model:

$$\left\{\mathbf{E}\left[(1-s)(Y_i^* + Z_{i,J+1}) + s\left(\frac{1}{2}\left(\frac{J}{2}+1\right) + \frac{1}{2}\left(\frac{J}{2}\right)\right)\right] - \mathbf{E}\left[(1-s)Y_i^* + s\frac{J}{2}\right]\right\} - \tau =$$
$$(1-s)\tau + s\left(\frac{J}{2}+\frac{1}{2}\right) - s\frac{J}{2} - \tau = s\left(\frac{1}{2}-\tau\right)$$

Also, the middle point for a control list with a placebo item and the treatment list will be the same, which again makes the bias when adding the placebo item $-s\tau$.

Hence, the consequences in terms of bias of the DiM estimator are the same for all tree error models (and the bias is constant across models when adding a placebo item). This also means that any weighted average of these three error models will exhibit the same bias. Note, however, that the *variance* will not be the same across the different models, given that (for instance) $Var(Binom(J+1, 0.5)) \neq Var(U\{0, J+1\})$ in general.

Forth, there is the possibility of a *top-biased* error model, under which respondent's true response is randomly replaced with the maximum value available: i.e. innatentives in the control group will chose J,

[13]The bias will be the same when J is uneven.

and $J+1$ when in treatment. As shown by Blair, Chou, and Imai (2019 appendix A), the bias of the DiM estimator under the top-biased error model will be:

$$\left\{ \mathbf{E}\left[(1-s)(Y_i^* + Z_{i,J+1}) + s(J+1)\right] - \mathbf{E}\left[(1-s)Y_i^* + sJ\right] \right\} - \tau = (1-s)\tau + s - \tau = s(1-\tau)$$

However, adding a placebo item (with expected value equal to 0) to the control list under the top-biased error model the bias becomes $-s\tau$. When does the addition of a placebo item decrease the amount of bias under the top-biased-error model? This happens when $s\,(1-\tau) > s\tau$. Hence:

$$s\,(1-\tau) > s\tau \Rightarrow 1-\tau > \tau \Rightarrow 1 > 2\tau \Rightarrow \tau < \frac{1}{2}$$

Under this error model the addition of a placebo item will therfore decrease the amount of bias as long as the true prevalence of $Z_{i,J+1}$ is below $\frac{1}{2}$.

Additional simulation evidence

For illustration, we simulate the inclusion of a placebo item on the control list when a share of respondents answer according to the uniform error model. Figure 2 displays the results from a simulation (using the same setup as above) that compares the $\hat{\tau}$ of a sample with 30% inattentive respondents without including a placebo item (purple) and 30% inattentive respondents including a placebo item (yellow). For the inattentive respondents in the control group with the placebo item the responses were hence drawn from $U\{0,5\}$, instead of $U\{0,4\}$. In this set up, using a control list without a placebo item substantively bias the estimate upwards (by $s\left(\frac{1}{2}-\tau\right)$ on average). The inclusion of the placebo item also bias the estimate (downwards, by $-s\tau$ on average), although less so. The existence and magnitude of these biases thus depend on the prevalence rate of the "sensitive item". The inclusion of the placebo item can thus improve the precision of the estimate, even if no inattentive respondents are excluded. In our simulation the inclusion of the placebo item produces a mean $\hat{\tau}$ of 0.12, just short of 5 percentage points away from the true prevalence of 0.16. This is less than half of the bias that we when estimating the prevalance without a placebo item (mean $\hat{\tau}$ 0.26.).

List effect bias

In list experiment B and E respondents are randomly assigned to one of three groups: a "normal" control group, a control group receiving the placebo item, and the treatment group receiving the "sensitive" item.

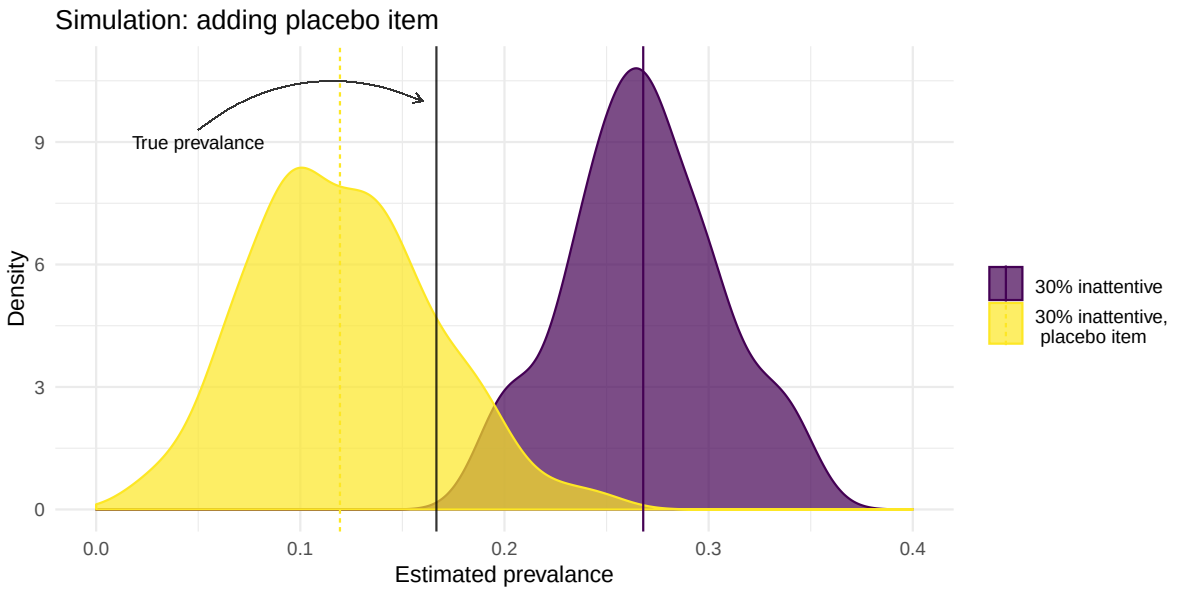

Figure 2: Simulation of $\hat{\tau}$ by adding a placebo item. 10000 simulated data sets.

We define list effect bias as the difference between any list effect in a list with J items and a list with $J+1$ items. We denote this bias ΔL. To test for the presence of ΔL we first exclude inattentive respondents and then compare the normal control list to the control list with a placebo item. In the absence of list effect bias this difference should be close to 0.

List effect bias implies that respondents given the longer list adjust their response in a way that that respondents assigned to the shorter list do not. To simulate this process we assume that respondents assigned to the normal control list respond according to their true preference for the control items denoted by Z_{ij}. Like in previous simulations we assume that each individual item is a draw from a Bernoulli distribution where the parameter p_j was set to 0.5, 0.5, 0.15, and 0.85 for the different items respectively. We assume that "respondents" in the placebo control group answer affirmatively to each item with the same probability, and never answer affirmatively to the placebo item. However, conditional on giving a "low" response these respondents adjust their answer upwards by one item (simulating positive list effect bias) with probability p_l. A response is "low" when the total number of affirmative answers is below som threshold j_{low}. For respondents in the normal control group the response is generated by: $Y_i^* = \sum_{j=1}^{J} Z_{ij}$. For respondents in the treatment group the response is generated by $Y_i^* = \sum_{j=1}^{J} Z_{ij} + \mathbb{1}_{\{\sum_{j=1}^{J} Z_{ij} < j_{low}\}} X_l$. Where $\mathbb{1}$ is the indicator function and $X_l \sim Bernoulli(p_l)$. Under this setup ΔL will be equal to $p(\sum_{j=1}^{J} Z_{ij} < j_{low}) \times X_l$.

To roughly estimate the power of our test to detect list effect bias (see above) we run the following simulation. We assume 4000 respondents. We assume that $s = 0.25$ and therefore exclude 1000 "inattentive" respondents from the sample, leaving 3000 respondents (hence around 1000 respondents in each group). We

set j_{low} to 3 and p_l to 0.1, indicating that respondents given the placebo list who would answer below 3 have a probability of 0.1 to adjust their answer upwards by one item. Under this setup the test comparing the difference between the two control lists (using the DiM estimator) has a 75% chance to reject the null hypothesis. Using the same simulation setup and a sensitive item with a prevalence of $\frac{1}{6}$ for the treatment group, we estimate that the chance that the $pRMSE(\hat{\tau})$ will be lower when estimating τ using the placebo control group as compared to the normal control group is 93%.