Three Essays on Survey Methods and Their Applications to Measuring Political Behavior and Attitudes

Thesis by
Yimeng Li

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended April 26, 2022

Yimeng Li
ORCID: 0000-0003-3855-0756

ACKNOWLEDGEMENTS

I'm grateful to my committee members, R. Michael Alvarez, Jonathan N. Katz, Alexander V. Hirsch, and Robert Sherman, for their encouragement, support, and advice. I'm indebted to my advisor R. Michael Alvarez for being a great mentor and collaborator, helping me succeed academically and professionally, and supporting me during challenging job market conditions. I express my gratitude to my thesis committee chair Jonathan N. Katz and option representative Alexander V. Hirsch for their advice on my job market paper, job applications, and career decisions, among other things.

I am also grateful to my best friend Jeffrey R. Zeidel for his company, support, and tolerance during the COVID-19 pandemic. It has been a challenging time both academically and personally, and thanks to him, I managed to keep my spirits up to the extent possible. I feel privileged to have a good friend like him.

For helpful conversations and comments on my research, I also thank other faculty members in and outside Caltech, including but not limited to Lonna Atkeson, Federico Echenique, Michael J. Ewens, Michael Gibilisco, Ines Levin, Gabriel Lopez-Moctezuma, Thomas R. Palfrey, III, and Matt Shum, and fellow graduate students and postdoctoral scholars, including but not limited to Jian Cao, Sumit Goel, Daniel R. Guth, Wade D. Hann-Caruthers, Seo-young Silvia Kim, Aldo Lucia, Lucas Núñez, Alejandro Robinson-Cortés, Jeffrey R. Zeidel, and Shiyu Zhang. I thank Kapáuhi Stibbard and Laurel Auchampaugh for administrative assistance and Daniel Yoder and Laura Flower Kim at the International Student Programs for their advice on immigration-related matters.

I thank the John Randolph Haynes and Dora Haynes Foundation for providing funding to support the research reported in Chapter 4 and the Davisson Fellowship and Repetto Fellowship for funding me in Fall 2018, Summer and Fall 2019, and Winter 2020. I also thank Neal Kelley and his team at the Orange County Registrar of Voters (OCROV) office for answering many of my questions and facilitating the transfer of the data regarding several research papers.

Finally, I am grateful to my parents for supporting me in my education and pursuing my research career. The journey would not have been possible without their support.

ABSTRACT

In this thesis, I develop survey methods and apply them to measure political behavior or attitudes more accurately. Two of the challenges to researchers in measuring political behavior or attitudes are respondents' reluctance to respond to sensitive questions truthfully and respondents' inattention to providing accurate responses. I contribute to advancing survey methodology to tackle these challenges.

Eliciting truthful answers from respondents on sensitive issues is a difficult problem in surveys, and list experiments emerge as the most popular indirect questioning technique to do so among political scientists and sociologists. The analysis of list experiments depends on two assumptions, known as "no design effect" and "no liars." The no liars assumption is strong and may fail in many list experiments. In Chapter II (published in *Political Analysis*), I relax the no liars assumption and develop a method to provide bounds for the prevalence of sensitive behaviors or attitudes under a weaker behavioral assumption. I apply the method to a list experiment on the anti-immigration attitudes of California residents and a broad set of existing list experiment datasets. My results indicate that the bounds tend to be narrower when the list consists of items of the same category, such as multiple groups or organizations, different corporate activities, and various considerations for politician decision-making. The contribution of my paper is to illustrate when the full power of the no liars assumption is most needed to pin down the prevalence of the sensitive behavior or attitudes, and to facilitate analysis of list experiments robust to violations of the no liars assumption.

Over the past two decades, the environment in which respondents participate in surveys and polls has changed, with shifts from interviewer-driven to respondent-driven surveying, and from probability to nonprobability sampling. One consequence of these technological changes is that survey respondents in these environments may be less attentive to survey questions. In Chapter III (published in *Political Analysis*), co-authored with R. Michael Alvarez, Lonna Atkeson, and Ines Levin, we study respondent attention and its implications using data from a self-completion online survey that identified inattentive respondents using instructed-response items (IRIs), a simple attention check that received little scholarly attention. Our results demonstrate that ignoring attentiveness provides a biased portrait of the distribution of critical political attitudes and behavior of both sensitive and more prosaic nature, and results in violations of key assumptions underlying experimental designs. We

discuss four approaches to dealing with inattentiveness in surveys and when these approaches are appropriate.

Attention checks, in the form of instructional manipulation checks (IMCs) or instructed response items (IRIs), are useful tools for survey quality control. However, due to the lack of ground truth information, these previous works rely on various post hoc measures to evaluate the performance of attention filters. For the same reason, it has also been impossible to evaluate the performance of different statistical approaches to dealing with inattentive respondents. In Chapter IV, co-authored with R. Michael Alvarez, we conduct a first validation study by analyzing a large-scale post-election survey following the November 2018 General Election and validating survey responses at the individual level using administrative records. Our results show that for each type of attention check, respondents failing the check provided responses with lower accuracy than respondents passing it. We compare the performance of different approaches to dealing with inattentive respondents in the study of turnout and voting method, two variables of substantive interest that are available from the administrative record, and conclude that the best strategy depends on a bias-variance trade-off that also accounts for the correlation between respondent attention and the outcome variables of interest.

PUBLISHED CONTENT AND CONTRIBUTIONS

Alvarez, R. Michael and Yimeng Li (2021). "Survey Attention and Self-Reported Political Behavior". en. In: *APSA Preprints*. DOI: 10.33774/apsa-2021-x689s. URL: https://preprints.apsanet.org/engage/apsa/article-details/61a901af704d057d023da5cf.

Y.L. participated in designing the research, reviewing the literature, developing the methodology, analyzing the data, and writing and revising the paper.

Alvarez, R. Michael, Lonna Rae Atkeson, et al. (2019). "Paying Attention to Inattentive Survey Respondents". en. In: *Political Analysis* 27.2, pp. 145–162. DOI: 10.1017/pan.2018.57. URL: https://www.cambridge.org/core/journals/political-analysis/article/paying-attention-to-inattentive-survey-respondents/BEDA4CF3245489645859E7E6B022E75A.

Y.L. participated in designing the research, reviewing the literature, developing the methodology, analyzing the data, and writing and revising the paper.

Li, Yimeng (2019). "Relaxing the No Liars Assumption in List Experiment Analyses". en. In: *Political Analysis* 27.4, pp. 540–555. DOI: 10.1017/pan.2019.7. URL: https://www.cambridge.org/core/journals/political-analysis/article/relaxing-the-no-liars-assumption-in-list-experiment-analyses/C0296899265E94123B30C5CBDF65B51B.

Y.L designed the research, reviewed the literature, developed the methodology, analyzed the data, and wrote and revised the paper.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

Many topics social scientists study are sensitive in nature, such as race-based discrimination, vote-buying in democracies, and support for leaders in authoritarian regimes. These are behavior or attitudes that are illegal, legal but dangerous, embarrassing, or against the social norm. Eliciting truthful answers from respondents on sensitive issues is a difficult problem in surveys. When surveyors ask respondents these questions directly, some respondents may not be willing to answer truthfully or may even refuse to answer. But often, researchers are not interested in determining whether a particular respondent has specific sensitive behavior or attitudes. In these cases, a list experiment is a popular technique that allows researchers to estimate the prevalence of sensitive behavior or attitudes.

In a list experiment, researchers randomly assign respondents to a control or treatment condition. Control respondents see a list of J items, and the question asks them how many they would respond to in the affirmative. Treated respondents see an otherwise identical question except for the addition of a sensitive item. Standard analysis of list experiments depends crucially on two assumptions. The first assumption, no design effect, states that the inclusion of the sensitive item does not affect respondents' latent answers to control items. The second assumption, no liars, states that respondents give truthful latent answers for the sensitive item.

Ideally, researchers would like all respondents to answer the list experiment question consistently and truthfully. However, the no liars assumption is strong and may fail in many circumstances. In Chapter II, I first illustrate which proportions of different types of respondents are identified in the absence of the no liars assumption. I proceed to develop a method to estimate bounds for the prevalence of sensitive behavior or attitudes under a weaker behavioral assumption. I then apply my method to a list experiment on the anti-immigration attitudes of California residents and a broad set of existing list experiment datasets in the literature. Finally, I discuss when the bounds under my relaxed liars assumption are likely to be narrow and illustrate when the full power of the no liars assumption is most needed to pin down the prevalence of the sensitive behavior or attitudes.

High-quality survey responses are essential for political science, where surveys are a primary tool for testing theories of political behavior or attitudes. Traditionally surveys were conducted through face-to-face or telephone interviewing, where the presence of an interviewer helps keep respondents focused. With high non-response and cost, in-person and telephone surveying are less prevalent, while online respondent-driven surveying is increasingly adopted. This recent shift from face-to-face and telephone surveying to respondent-driven online surveying has introduced challenges in keeping respondents attentive and providing quality survey responses that reflect their actual behavior or attitudes. What are the implications for data quality? How can we identify inattentive survey respondents? What should we do to deal with survey inattentiveness?

A strategy for identifying inattentive survey respondents involves embedding attention checks in carefully selected locations in the survey instrument. One type of attention check is an instructional manipulation check (IMC), where there is a deliberate change in the instructions in a survey question designed to capture whether the respondent is reading and cognitively processing the question's instructions. An example of an IMC is adding a clause to a survey question instructing the respondent to ignore the question and provide a specific answer. In Chapter III, co-authored with R. Michael Alvarez, Lonna Atkeson, and Ines Levin, we measure respondents' attentiveness using another type of attention check called instructed-response items (IRIs), where the responses to a survey question are altered in a way that should elicit whether the respondent is attentive to the question's response options. An example of an IRI is adding a row in a grid instructing respondents to select 'strongly disagree' for survey quality control.

Using a 2014 online survey of California adults with multiple instructed-response items, we show that a substantial proportion of respondents pay little attention to survey questions. Younger and less educated respondents, in particular, are more likely to fail the attention checks. We also demonstrate how the responses of the inattentive respondents to important direct survey questions (political participation, political knowledge, political attitudes) are very different from the responses of the attentive respondents, and ignoring respondent attentiveness may lead to a biased evaluation of the incidence of critical attitudes and behaviors. We further document that inattentive respondents give noisy (and perhaps not very informative) responses to indirect survey questions about sensitive behaviors, which could challenge fundamental assumptions underlying the experimental design.

Due to the lack of ground truth information, previous work relies on various post-hoc measures to evaluate the performance of attention filters. These measures look at whether respondents passing and failing attention filters differ in producing canonical experimental results, empirical correlations between negatively correlated survey questions, straight-lining behavior, and response time. For the same reason, it has also been impossible to evaluate the performance of different statistical approaches to dealing with inattentive respondents. In Chapter IV, co-authored with R. Michael Alvarez, we use unique data from 2018 to conduct the first validation study by looking at responses to factual survey questions that we can validate with external administrative data.

We find that respondents failing the attention checks are more likely to misreport information than those passing the attention checks, and including inattentive respondents as measured by the attention checks will lead to bias in the estimates. Many respondents failing the attention checks nonetheless provide accurate self-reports, implying that dropping inattentive respondents as measured by the attention checks will increase the variance in the estimates. Our results also indicate that if respondent attention is correlated with the outcome of interest, dropping inattentive respondents can produce an unrepresentative sample and thus bias estimates. Therefore, the best strategy to deal with inattentive respondents depends on the bias-variance trade-off and the correlation between respondent attention and the outcome of interest. We discuss the trade-off between different approaches in dealing with inattentive respondents.

Chapter 2

RELAXING THE NO LIARS ASSUMPTION IN LIST EXPERIMENT ANALYSES

Li, Yimeng (2019). "Relaxing the No Liars Assumption in List Experiment Analyses". en. In: *Political Analysis* 27.4, pp. 540–555. DOI: 10.1017/pan.2019.7. URL: https://www.cambridge.org/core/journals/political-analysis/article/relaxing-the-no-liars-assumption-in-list-experiment-analyses/C0296899265E94123B30C5CBDF65B51B.

2.1 Introduction

Eliciting truthful answers from respondents on sensitive issues is a difficult problem in surveys. When surveyors ask respondents these questions directly, some respondents may not be willing to answer truthfully or may even refuse to answer. To address this problem, several survey techniques have been developed and have gained popularity among political scientists and sociologists, such as list experiments (item count technique), randomized response techniques, and endorsement experiments. In political science, scholars have conducted and analyzed list experiments on various topics involving sensitive behaviors or attitudes (early works include Kuklinski, Cobb, and Gilens, 1997 on racial prejudice; Kane, S. C. Craig, and Wald, 2004 on religion; Streb et al., 2008 on gender; Corstange, 2009 on turnout; and Holbrook and Jon A. Krosnick, 2010 on voting rights; see Section 2.3 for a list of recent works).

List experiments are intended to eliminate the direct association between a respondent's answer and his or her attitude toward the sensitive item by including a few nonsensitive items ("control items").[1] By doing so, list experiments give respondents the opportunity to provide truthful answers to the question while not revealing their attitudes toward the sensitive issue. Even though researchers do not gain knowledge about individual respondents' preferences toward the sensitive issue, researchers can obtain the proportion of the population whose truthful responses are affirmative to the sensitive issue under some important assumptions. In their

[1] I describe the structure of list experiments in detail in section 2.2. In a nutshell, respondents in the control group see a list of control items, and respondents in the treatment group see a list of the sensitive item and control items. Respondents report how many items they would respond to in the affirmative.

research on list experiments, Imai, 2011 and Blair and Imai, 2012 formulate the following two assumptions: (1) the inclusion of the sensitive item has no effect on respondents' latent answers to the control items; (2) respondents give truthful latent answers to the sensitive item.[2] The analysis of list experiments depends crucially on these two assumptions, with the first assumption called "no design effect" and the latter referred to as "no liars". The standard estimator is the difference between the average response from treated respondents and that from control respondents. The estimator is simple, intuitive, and unbiased under the assumptions of no design effect and no liars.

The no liars assumption is strong and may fail in many list experiments. In particular, a treated respondent who favors all or a large number of control items and the sensitive item may be reluctant to answer truthfully, as it reveals with certainty or a high probability that the respondent favors the sensitive item. In a recent study, Rosenfeld, Imai, and Shapiro, 2016 finds that their list experiment underestimates high incidence behavior in the context of a telephone interview. One possible explanation for this observation is that some respondents give untruthful answers for the sensitive item. Statistical tests are useful to detect certain forms of violations, but they usually look at the no design effect assumption, may lack power, or only examine a small subset of respondents. In this paper, I relax the no liars assumption, and in order to do so, maintain the no design effect assumption and focus on behaviors or attitudes with one-sided sensitivity.

I first illustrate which quantities concerning the proportions of different types of respondents are identifiable in the absence of the no liars assumption. In particular, the proportions of truth-telling respondents with the sensitive behaviors or attitudes are identifiable, but liars are observationally equivalent to respondents who do not favor the sensitive item, conditional on the same number of control items answered affirmatively. I proceed to develop a method to estimate bounds for the prevalence of sensitive behaviors or attitudes that does not depend on the no liars assumption. To relax the no liars assumption, I introduce parameters that capture the proportion of liars conditional on the number of control items answered affirmatively. Under a mild condition that the respondents who are supposed to answer affirmatively to all control items as well as the sensitive item have the strongest incentive to lie, I can derive bounds for the level of support for the sensitive item. In particular, the lower bound is the standard difference-in-means estimates. For the upper bound, I first

[2] Under the first assumption (no design effect), a respondent's answer is the sum of their latent answers to the control items and the sensitive item.

calculate the maximal ratio of liars versus truth-tellers among respondents answering affirmatively to all control items, and use this ratio to bound the proportion of liars whose answer is affirmative to fewer than all control items. To permit inference, I construct the confidence set using techniques developed for partially identified models (Imbens and Manski, 2004; Stoye, 2009).

I then apply my method to a list experiment on the anti-immigration attitudes of California residents and on a broad set of existing list experiment datasets in the literature. In my example on anti-immigration attitudes, I show that I can reach substantive conclusions without having to put full faith on the no liars assumption. I discuss when the bounds under my relaxed liars assumption are likely to be narrow. In particular, my simulated list experiments suggest that positive correlations between the sensitive item and the control items implies narrower bounds. Since such positive correlations are likely to be present when items on a list are of the same category, this result is consistent with my finding of narrower bounds in list experiments conducted in the literature with this feature. The contribution of my paper is to illustrate when the full power of the no liars assumption is most needed to pin down the prevalence of the sensitive behavior or attitude, and to facilitate analysis of list experiments robust to violations of the no liars assumption.

The structure of the rest of the paper is as follows. I formally describe list experiments and introduce my method in Section 2.2. In Section 2.3, I illustrate my method via an example and apply my method to a broad set of list experiment datasets in published studies as well as some simulated list experiments. I discuss a stronger assumption and potential challenges in Section 2.4 and then conclude.

2.2 Method

Setup

Consider a list experiment with N respondents, randomly sampled from the population of interest. Researchers randomly assign the respondents to a control or a treatment condition. The number of respondents under control and treatment need not be the same. Researchers present to the control respondents a list of J control items and ask them how many they would respond to in the affirmative. For treated respondents, researchers ask an otherwise identical question except for the addition of a sensitive item, namely among a list of J control items and a sensitive item, how many they would respond affirmatively to.

The canonical difference-in-means estimator is the difference between the average of respondents' answers under treatment and control. Imai, 2011 and Blair and Imai, 2012 formulate the following two identification assumptions that facilitate the analysis of list experiments:

> *Assumption 1 (No design effect)*: The inclusion of the sensitive item has no effect on respondents' answers to control items,

and

> *Assumption 2 (No liars)*: Respondents give truthful answers for the sensitive item.

Under Assumption 1, the difference in means is simple and intuitive, and it gives an unbiased estimator of the population proportion of those who give an affirmative answer to the sensitive item. Under Assumption 2, it is also the population proportion of those who favor the sensitive item in their underlying preferences.

I make three remarks here. First, following Imai, 2011 and Blair and Imai, 2012, researchers do not assume respondents give truthful answers to the control items, only that they answer in a consistent way under control and treatment. Second, I make a distinction between a respondent giving an affirmative answer to the sensitive question and the respondent favoring the sensitive item in their underlying preference. This distinction will be important throughout the paper. Finally, in the main text, I describe my method for the case where an affirmative response to the sensitive item is considered sensitive (such as have done, support, or agree with the sensitive item). The case where a negative response is sensitive is analogous and described in Section A.5 in the supplementary materials.

An important implication of no design effect and no liars is that researchers can recover—from the distribution of responses under control and treatment—the joint distribution of respondents' preferences toward the sensitive item and the number of control items answered affirmatively (Glynn, 2010). In particular, denote by p_k the proportion of the population who respond affirmatively to k control items and do not favor the sensitive item, and by p_{kT} the proportion responding affirmatively to k control items and favoring the sensitive item (shown in Table 2.1). Then $(p_k, p_{kT})_{k=0}^{J}$ are identified under Assumption 1 and 2.[3]

[3] To see this, start with p_{JT}. Since under Assumption 1 and 2 respondents who answer $J+1$

# of Control Items	Sensitive Item No	Sensitive Item Yes
0	p_0	p_{0T}
1	p_1	p_{1T}
2	p_2	p_{2T}
3	p_3	p_{3T}
4	p_4	p_{4T}

Table 2.1: Distribution of different types of respondents ($J = 4$)

No Liars Assumption

Ideally researchers would like all respondents to answer the list experiment question in a consistent and truthful way, satisfying the no design effect assumption and the no liars assumption. However, the no liars assumption is strong and may fail in many circumstances. For a treated respondent whose answer to all control items is affirmative and who favors the sensitive item, truth telling reveals with certainty that the respondent's answer toward the sensitive item is affirmative, often referred to as ceiling effects in the literature (Kuklinski, Cobb, and Gilens, 1997; Blair and Imai, 2012; Glynn, 2013).[4] And for a treated respondent whose answer to a large number of control items is affirmative and who favors the sensitive item, intuitively, truth telling reveals with a high probability that the respondent's answer toward the sensitive item is affirmative. Respondents who feel strongly about the social norm and fear their preferences for the sensitive item might be revealed thus have an incentive to lie about the sensitive item under list experiments, which leads to a violation of the no liars assumption. Moreover, failing to understand the protective nature of the list experiment design, actively managing perceptions about oneself (Köszegi, 2006) and being too embarrassed to admit also lead respondents to form untruthful latent response to the sensitive item.

Statistical tests are useful to detect certain violations of the assumptions, but there are limitations to these tests. Blair and Imai, 2012 propose to test the positivity of the proportions of different types of respondents, $(p_k, p_{kT})_{k=0}^{J}$. However, they

under treatment consist exclusively of those who answer affirmatively to k control items and favor the sensitive item, one can estimate p_{JT} by the proportion giving the maximal answer under treatment. Now respondents who answer J under control consist of those who answer affirmatively to k control items, regardless of their preferences to the sensitive item. So one can estimate p_J by subtracting the estimated p_{JT} from the proportion giving the maximal answer under control. One can continue in this fashion and obtain all $(p_k, p_{kT})_{k=0}^{J}$.

[4] Similarly, for a treated respondent whose answer to all items—control and sensitive—is negative, truth telling is also fully revealing, known as floor effects.

note that "the violation of Assumption 2 [no liars] alone does not lead to negative proportions of these types" and "researchers may fail to reject the null hypothesis due to a lack of statistical power" (Blair and Imai, 2012, p. 64–65). In other words, the utility of the test is primarily detecting failures of the no design effect assumption and lack of statistical power can be a concern. Follow-up work by Aronow et al., 2015 proposes a complementary test when a direct question to the same respondents is also available.[5] Their test, however, can only detect violations of list experiment assumptions for respondents who reveal preferences for the sensitive item under the direct question.

In a recent validation study, Rosenfeld, Imai, and Shapiro, 2016 find their list experiment underestimates sensitive votes against a referendum. One possible explanation for this observation is that some respondents give untruthful answers for the sensitive item.

In this paper, I will maintain the no design effect assumption and relax the no liars assumption. It will become clear that the no design effect assumption has important implications for identifying different types of respondents on its own and my analysis depends critically on it. For further discussion of this assumption, see Section A.6 in the supplementary materials.

Identification in the Absence of No Liars Assumption

In the absence of the no liars assumption, the decomposition of respondent types in Table 2.1 is no longer valid. Instead, I denote the population fraction of respondents who answer affirmatively to k control items and (1) do *not* favor the sensitive item by p_{kN}; (2) favor the sensitive item, but would *not* give the truthful answer for it by p_{kL}; (3) favor the sensitive item, and give the truthful answer for it by p_{kT}. In particular, the proportion p_k of respondents who answer in the affirmative to k control items now consists of both the proportion p_{kN} of non-supporters and the proportion p_{kL} of liars. Assuming only one side of the sensitive item is potentially socially undesirable, i.e.,

> *One-sided sensitivity assumption*: Respondents who do not favor the sensitive item do not lie and falsely give an affirmative answer.

[5] Direct questions also set up useful benchmarks for assessing how well list experiments reduces social desirability bias (Blair and Imai, 2012; Kramon and Weghorst, 2012).

I can present the population distribution of different types of respondents in Table 2.2. In this terminology, the no liars assumption states that $p_{kL} = 0$ for all k.

# of Control Items	Sensitive Item Non-supporters	Liars	Truth-tellers
0	p_{0N}	p_{0L}	p_{0T}
1	p_{1N}	p_{1L}	p_{1T}
2	p_{2N}	p_{2L}	p_{2T}
3	p_{3N}	p_{3L}	p_{3T}
4	p_{4N}	p_{4L}	p_{4T}

Table 2.2: Distribution of different types of respondents ($J = 4$)

I do not observe a respondent's latent answer for the control items or the latent attitude toward the sensitive item (except when the respondent gives an answer of 0 or J). What I observe is a response from 0 to J for each respondent under the control condition, and a response from 0 to $J + 1$ for each respondent under the treatment condition. Let c_k be the population proportion of respondents whose answer is k under the control condition and t_k be the analogous proportion under the treatment condition, with $\sum_{k=0}^{J} c_k = \sum_{k=0}^{J+1} t_k = 1$.

To map a respondent's latent response toward the control items and the latent attitude toward the sensitive item to his or her answer under the control condition, take a respondent who answers one item as an example. This respondent must answer affirmatively to exactly one control item, but I don't have any information about his or her attitude toward the sensitive item. The respondent may or may not favor the sensitive item, and in the latter case might or might not give a truthful answer for it if he or she were assigned to the treatment condition instead. Therefore, I have $c_1 = p_{1N} + p_{1L} + p_{1T}$. In the same fashion, I can express each c_k in terms of the p-terms, as shown in the left column in Table 2.3.

For a treated respondent, I can establish the mapping in a similar way. For example, to see $t_1 = p_{1N} + p_{1L} + p_{0T}$, notice that a response of one item under treatment comes from a respondent who (1) answers affirmatively to exactly one control item and does not favor the sensitive item, or (2) answers affirmatively to exactly one control item and favors the sensitive item, but does not give a truthful answer for it, or (3) answers affirmatively to no control items, but favors the sensitive item and is truthful. The expressions for all t_k in terms of the p-terms are shown in the right column in Table 2.3.

$c_0 = p_{0N} + p_{0L} + p_{0T}$	$t_0 = p_{0N} + p_{0L}$
$c_1 = p_{1N} + p_{1L} + p_{1T}$	$t_1 = p_{1N} + p_{1L} + p_{0T}$
$c_2 = p_{2N} + p_{2L} + p_{2T}$	$t_2 = p_{2N} + p_{2L} + p_{1T}$
$c_3 = p_{3N} + p_{3L} + p_{3T}$	$t_3 = p_{3N} + p_{3L} + p_{2T}$
$c_4 = p_{4N} + p_{4L} + p_{4T}$	$t_4 = p_{4N} + p_{4L} + p_{3T}$
	$t_5 = p_{4T}$

Table 2.3: Relationship between latent attitudes and observable responses ($J = 4$)

In Table 2.3, there are more unknown parameters than equations. Hence, the system of equations is under-identified and I cannot identify the distribution of all latent types, $(p_{kL}, p_{kN}, p_{kT})_{k=0}^{J}$. A closer inspection reveals the reason why this system is under-identified. A lying respondent who actually favors the sensitive item is observationally equivalent to a respondent who does not favor the sensitive item at all, conditional on the same number of control items answered affirmatively. As a result, only the sum of the proportions of these two types is identified. In particular, regarding $p_{kN} + p_{kL}$ as one parameter instead of two parameters p_{kN} and p_{kL}, the system is exactly identified and I can identify $p_{kN}+p_{kL}$ and p_{kT} for each $k = 1, ..., J$.

Relaxed Liars Assumption, Bounds and Inference

The quantity of interest is the population proportion of respondents who favor the sensitive item, $\sum_{k=0}^{J} p_{kL} + p_{kT}$. I have already shown that the proportion of truth-tellers who respond affirmatively to k control items, p_{kT}, is identified for each k, the remaining work concerns the fraction of respondents who favor the sensitive item but do not give truthful answers for it, $\sum_{k=0}^{J} p_{kL}$. Since $p_{kN} + p_{kL}$ is identified for each $k = 1, ..., J$, by setting all p_{kL}'s to be zero (no liars), I obtain the standard difference-in-means estimator as a lower bound of the level of support for the sensitive item. At the other extreme, if I set all p_{kN}'s to be zero (no non-supporters), I obtain a trivial upper bound of 1.

With no behavioral assumption about the pattern of lying, I do not have an informative upper bound on the prevalence of the sensitive item. Toward a non-trivial upper bound, I propose the following assumption concerning the incentive to lie:

Relaxed liars assumption:

$$\frac{p_{kL}}{p_{kL} + p_{kT}} \leq \frac{p_{JL}}{p_{JL} + p_{JT}}, \ \forall k = 0, \ldots, J-1, \tag{2.1}$$

where $p_{kL}/(p_{kL} + p_{kT})$ is the ratio of liars conditional on responding affirmatively to k control items and favoring the sensitive item. This assumption says that among all respondents who favor the sensitive item, the ones who respond affirmatively to all control items have the strongest incentive to lie. It relaxes the assumption of no liars, which amounts to $p_{kL}/(p_{kL}+p_{kT}) \equiv 0$ for all $k=0,\ldots,J$. The relaxed liars assumption is simple and intuitive, and it captures the intuition behind the ceiling effect that truth telling reveals with certainty a respondent's preference for the sensitive item when he or she favors all items. Also note that this assumption does not impose any relationship between lying incentives for respondents favoring fewer than all control items.

My relaxed liars assumption gives an upper bound for the number of liars among respondents who favor fewer than all control items. However, I need an estimate or upper bound on p_{JL} to make this assumption implementable. Since $p_{JN}+p_{JL}$ is identified, it can serve as an upper bound. This bound will be crude unless p_{JN} is relatively small. As presented in Section 2.3, in some applications, the crude bound can nevertheless be informative. In other list experiments, the bound can be wide, in which case the no liars assumption is important to nail down the prevalence of the sensitive item. Nonzero lower bound on p_{kN}, $k=0,\ldots,J$, would improve upon the crude bounds, but would also necessarily involve additional information and/or assumptions. Formally, an upper/lower bound for the level of support for the sensitive item is given by the solution to the following linear system:

$$\begin{aligned}
&\max/\min_{(p_{kL},p_{kN},p_{kT})_{k=0}^{J}} \textstyle\sum_{k=0}^{J} p_{kL}+p_{kT}\\
&\text{s.t. } c_0 = p_{0N}+p_{0L}+p_{0T},\ \ldots,\ c_J = p_{JN}+p_{JL}+p_{JT}\\
&\quad t_0 = p_{0N}+p_{0L},\ \ldots,\ t_J = p_{JN}+p_{JL}+p_{J-1,T},\ t_{J+1} = p_{JT}\\
&\quad p_{kL}/(p_{kL}+p_{kT}) \le (p_{JN}+p_{JL})/(p_{JN}+p_{JL}+p_{JT}),\ \forall k=0,\ldots,J-1.
\end{aligned} \tag{2.2}$$

I derive the bounds in steps in Section A.1 in the supplementary materials, and the lower and upper bounds are given by

$$\Big[\sum_{k=0}^{J} p_{kT}, \sum_{k=0}^{J} p_{kT} + \sum_{k=0}^{J} \min\{\frac{\lambda}{1-\lambda} p_{kT}, p_{kN}+p_{kL}\}\Big], \tag{2.3}$$

where $p_{kT} = \sum_{i=k+1}^{J+1} t_i - \sum_{j=k+1}^{J} c_j$, $p_{kN}+p_{kL} = \sum_{j=k}^{J} c_j - \sum_{i=k+1}^{J+1} t_i$, and the maximal liar ratio $\lambda \equiv (p_{JN}+p_{JL})/(p_{JN}+p_{JL}+p_{JT}) = (c_J - t_{J+1})/c_J$. It follows that the bounds are nontrivial if and only if $\frac{\lambda}{1-\lambda} p_{kT} \le p_{kN}+p_{kL}$, or equivalently $p_{kT}/(p_{kN}+p_{kL}) \le p_{JT}/(p_{JN}+p_{JL})$ for some k. In words, the condition holds if the proportion of truth-telling supporters is higher for respondents answering affirmatively to all J control items than $k < J$ control items. A higher prevalence of the sensitive item will increase $\{p_{kT}\}_{k=0}^{J}$ but decrease λ, so the overall effect is ambiguous. Factors holding the prevalence fixed while decreasing the maximal liar ratio will narrow the bounds. I conduct detailed investigations in Section 2.3.

After identifying the lower and upper bounds, which I denote by $\hat{p}_l$ and $\hat{p}_u$, I can construct the $(1-\alpha)$-confidence intervals $\mathrm{CI}(\hat{p}_l)$ and $\mathrm{CI}(\hat{p}_u)$ via boostrap. The confidence set formed by the lower endpoint of $\mathrm{CI}(\hat{p}_l)$ and the upper endpoint of $\mathrm{CI}(\hat{p}_u)$ covers the identified set with probability $(1-\alpha)$, but will be slightly conservative in terms of covering the true prevalence of the sensitive item. Alternatively, I can construct the confidence set with $(1-\alpha)$-coverage following Imbens and Manski, 2004 and Stoye, 2009 as

$$\mathrm{CI}_\alpha = \left[\hat{p}_l - c_\alpha \frac{\hat{\sigma}_l}{\sqrt{N}}, \hat{p}_u + c_\alpha \frac{\hat{\sigma}_u}{\sqrt{N}}\right], \tag{2.4}$$

where N is the total sample size, $\hat{\sigma}_l$ ($\hat{\sigma}_u$) is the standard error for $\hat{p}_l$ ($\hat{p}_u$), and c_α solves

$$\Phi\left(c_\alpha + \frac{\sqrt{N}}{\max\{\hat{\sigma}_l, \hat{\sigma}_u\}}(\hat{p}_u - \hat{p}_l)\right) - \Phi(-c_\alpha) = 1 - \alpha, \tag{2.5}$$

where Φ is the standard normal cumulative distribution function.[6,7] For a 95% confidence set, c_α will be $\Phi^{-1}(0.975) \approx 1.96$ if $\hat{p}_u - \hat{p}_l = 0$ and will approach $\Phi^{-1}(0.975) \approx 1.64$ as $\hat{p}_u - \hat{p}_l$ grows large relative to sampling error.

Extensions

Auxiliary Information to Sharpen Bounds

Further information can help us bound p_{kN} and improve upon the crude upper bound. Auxiliary direct questions can be helpful to sharpen the bound. For example, researchers interested in learning the prevalence of anti-immigration attitudes can ask a direct question on support for a pathway to citizenship with possible choices yes, no, and ambivalent answers (not sure, rather not say, or don't know). The proportion of yes answers—indicating pro-immigration attitudes—among those who favor k control items as a conservative estimate for p_{kN}, under the assumption that the

[6] Notice that the upper bound is continuous but not differentiable at points satisfying $p_{kT} \cdot \lambda/(1-\lambda) = p_{kN} + p_{kL}$. Outside this set of measure zero, the upper bound is differentiable and the delta method gives asymptotic normality pointwise. Moreover, outside neighborhoods of these nondifferentiability points (that can be taken to have arbitrarily small measure), standard argument gives uniform convergence to normal distribution asymptotically.

[7] In principle, one can calculate the asymptotic variance of $\hat{p}_l$ and $\hat{p}_u$, and obtain $\hat{\sigma}_l$ and $\hat{\sigma}_u$ by taking sample analogues. I avoid the cumbersome derivation by bootstrapping.

proportion of liars who answer yes to this direct question is smaller than that of non-supporters who give a negative or ambivalent answers.

Too Few Respondents on the Boundary

To reduce the number of respondents choosing zero or all items, some applied researchers include an item of very high prevalence and an item of very low prevalence. Glynn, 2013 instead suggests including negatively correlated items in favor of using high- or low-prevalence items. Both practices may result in very few respondents favoring all control items in some list experiments. In this case, the estimation of maximal liar ratio $(p_{JN} + p_{JL})/(p_{JN} + p_{JL} + p_{JT})$ in (2.2) relies on a small number of observations.

If answering affirmatively to all control items is vanishingly unlikely, then answering affirmatively to all but one control items is effectively an extreme response. So instead of estimating the maximal liar ratio as $(p_{JN} + p_{JL})/(p_{JN} + p_{JL} + p_{JT})$ in (2.2), I collapse the case of all control items and all but one control items and replace it with $(p_{J-1,N} + p_{J-1,L} + p_{JN} + p_{JL})/(p_{J-1,N} + p_{J-1,L} + p_{J-1,T} + p_{JN} + p_{JL} + p_{JT})$, a quantity that can be more precisely estimated.[8] In words, I now require respondents favoring at least all but one control items to have stronger incentive to lie than those favoring fewer items.

2.3 Applications

Illustrative Example

For illustration, consider a list experiment R Michael Alvarez et al., 2019 included in an online survey of 2,722 California adults conducted in 2014 using a sample recruited by Qualtrics. I excluded inattentive respondents who fail one of three trap questions included in the survey, and this leaves me with a total of 1,750 respondents.[9] The experiment was designed to measure the support for two state chapters of national anti-immigrant organizations among California residents. To

[8] When there are "too few" respondents answering the maximal number of items and to collapse the two cases is a subjective decision to make. For applications in Section 2.3, most list experiments have a sample size of 1000–2000 with a few notable exceptions. I decide to collapse the two cases whenever there are fewer than 50 respondents choosing all items (or zero items for list experiment with negative sensitive responses).

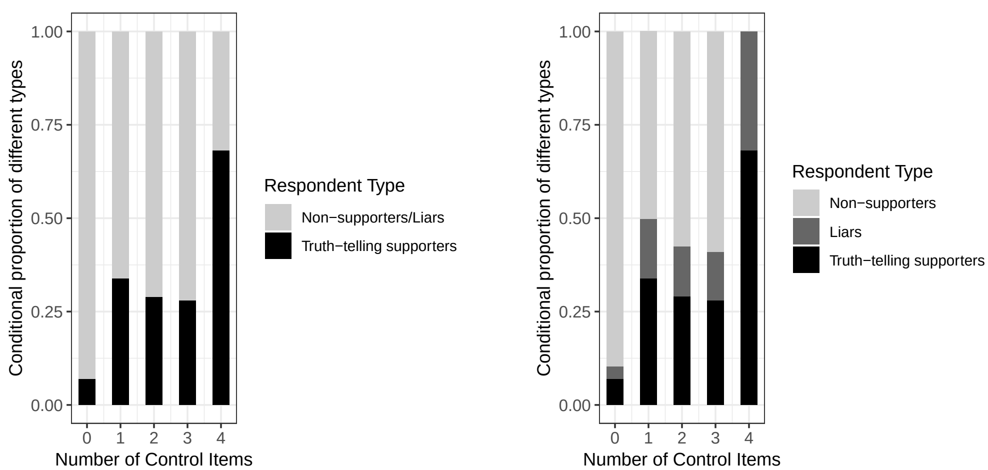
[9] R Michael Alvarez et al., 2019 implemented a double list experiment in fixed list order but for my purposes, I focus on the first list in this paper. I refer interested readers to R Michael Alvarez et al., 2019 for details on these trap questions and the differences in the response pattern between respondents who passed and failed the trap questions, in terms of their reported political knowledge, political participation, ideology, and responses to the list experiment.

preserve the anonymity of these organizations, I refer to them here as organization X and organization Y.

Respondents saw a control list or a treatment list with the names and short descriptions of different groups and organizations as shown in Table A.2, and were instructed to specify "how many of these groups and organizations you broadly support."

The proportions of control and treated respondents giving each answer are shown in Table A.3. Under the no design effect assumption, I can identify the proportions of truth-telling supporters and (truth-telling) non-supporters/liars conditional on the number of control items answered affirmatively. Figure 2.1 shows these proportions for organization X on the left. If I further impose the relaxed liars assumption and allow a maximal number of liars, I can obtain the proportion of truth-telling supporters, the maximal proportion of liars, and the minimal proportion of non-supporters, all conditional on the number of control items answered affirmatively. These proportions are shown in Figure 2.1 on the right.

Figure 2.1: Proportions of different types of respondents for Organization X



Note: The figure presents, for Organization X (the sensitive item), the proportions of truth-telling supporters and non-supporters/liars (left), and the proportions of truth-telling supporters, liars, and non-supporters under the relaxed liars assumption with a maximal number of liars (right), conditional on the number of control items answered affirmatively.

Doing the calculations, the difference-in-means estimates are 0.36 for Organization X and 0.22 for Organization Y. 95% confident intervals for the point estimates are $[0.19, 0.53]$ and $[0.05, 0.39]$, respectively. From my argument in Section 2.2, the difference-in-means estimate may understate the level of support for these organizations by neglecting respondents who support them but did not give truthful answers.

If I allow the possibility of liars, but assume my relaxed liars assumption and implement my crude bound, I obtain an upper bound 0.53 for the level of support for Organization X, and 0.36 for Organization Y. 95% confidence sets for the interval estimates are $[0.23, 0.80]$ and $[0.10, 0.64]$ for Organization X and Organization Y, respectively. My bounds estimates have important implications. To reach the conclusion that at most around a half respondents support Organization X and around a third respondents support Organization Y, I do not need to have full faith on the no liars assumption.

List Experiments in Published Studies

I now turn attention to a comprehensive set of list experiments in the political science literature, with basic information about these studies summarized in Table A.5.[10]

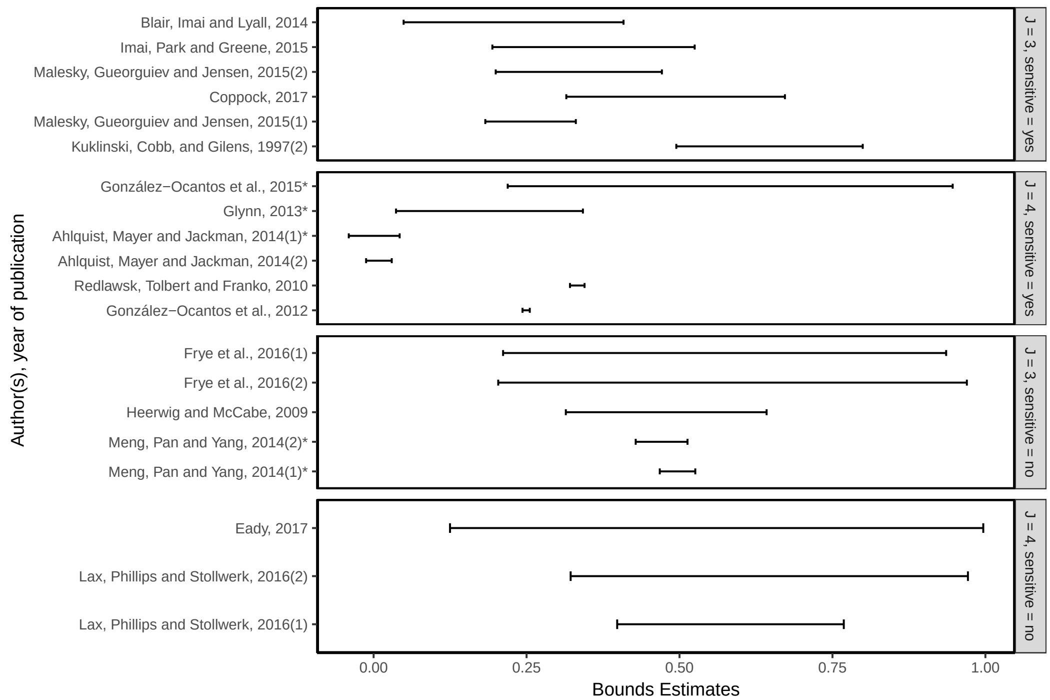
First notice that the list experiment concerning racial prejudice in Kuklinski, Cobb, and Gilens, 1997 and the list experiment concerning vote for an anti-abortion referendum in Rosenfeld, Imai, and Shapiro, 2016 fail the Blair-Imai test of no design effect. Therefore, I exclude these applications from subsequent analysis. I now apply my method to obtain the range of the prevalence of sensitive behaviors or attitudes possible under the relaxed liars assumption for all other list experiments. Since no extra information for constructing tighter bounds for these studies is available, I compute my crude bounds. Recall that the lower bounds correspond to the difference-in-means estimates under the no liars assumption. In Figure 2.2, the intervals show the estimated lower and upper bounds.[11] The top two panels and the bottom two panels display studies with affirmative and negative sensitive responses, respectively. Within each panel, the studies are ordered by the maximal proportion of liars permitted by the relaxed liars assumption (with the largest value on the top).

The widths of the intervals vary substantially across studies. Within each panel, the intervals tend to be narrower for studies with a smaller maximal proportion of liars permitted by the relaxed liars assumption. Across panels, the intervals appear to be narrower for list experiments with positive sensitive responses than those with

[10] I include—to the best my knowledge—all published list experiments conducted in political science with available replication data or the distribution of responses in the paper before 2017, with the exception of Aronow et al., 2015, C. P. Kiewiet de Jonge, 2015 and K. B. Coffman, L. C. Coffman, and Ericson, 2017 as they include multiple list experiments ($\geq$ 8) and I want to put together list experiments from a wide range of surveys and topics.

[11] 95% confidence sets for these bounds are suppressed to keep the figure uncluttered. As is the case for my illustrative example, the confidence sets are wide in general. The notable exceptions are Malesky, Gueorguiev, and Jensen, 2015, Coppock, 2017, and Eady, 2017 as these list experiments are implemented in larger scale.

Figure 2.2: Bound Estimates for List Experiments in Published Studies



Note: The figure displays the bound estimates for list experiments in Kuklinski, Cobb, and Gilens, 1997, Heerwig and McCabe, 2009, Redlawsk, Tolbert, and Franko, 2010, González-Ocantos, Jonge, et al., 2012, Glynn, 2013, Ahlquist, Mayer, and Jackman, 2014, Blair, Imai, and Lyall, 2014, Meng, Pan, and Yang, 2014, González-Ocantos, C. Kiewiet de Jonge, and Nickerson, 2015, Imai, Park, and Greene, 2015, Malesky, Gueorguiev, and Jensen, 2015, Frye et al., 2016, Lax, Phillips, and Stollwerk, 2016, Coppock, 2017, and Eady, 2017. For list experiments marked with stars, procedures described in Section 2.2 are adopted.

negative sensitive responses, with the exception of González-Ocantos, C. Kiewiet de Jonge, and Nickerson, 2015 and Meng, Pan, and Yang, 2014. However, this pattern can be due to difference in other aspects of these list experiments; in fact, there is a symmetry between the case of an affirmative sensitive item and a negative sensitive item.[12] To better understand why some intervals are narrow and some are wide, I take a closer look at these list experiments.

As I showed in Section 2.2 and illustrated in the example, under no design effect, I can identify the proportions of (1) truth-tellers with the sensitive behavior or attitude p_{kT}, and (2) the combination of liars p_{kL} and truth-tellers without the sensitive behavior or attitude p_{kN}, conditional on the number of control items answered

[12]Simulation results described in the next section also suggest whether the sensitive response is affirmative or negative does not drive the bound widths by itself.

affirmatively (or negatively for list experiments with negative sensitive responses). Figure A.1 and Figure A.2 display these conditional proportions for studies with affirmative and negative sensitive responses, respectively. In other words, the lines in these figures correspond to the line connecting the top of the black bars shown in Figure 2.1. Since crude bounds use no external information about liars versus non-supporters for sensitive items, narrower bounds come from: first, increasing proportions of truth-tellers with the sensitive behavior/attitude conditional on the number of control items answered affirmatively/negatively, which corresponds to the inequalities in the relaxed liars assumption (2.1); second, a smaller maximal proportion of liars permitted by the relaxed liars assumption, which corresponds to the right hand side of (2.1). As shown in Figure A.1 and Figure A.2, the increasing trend is present for many list experiments with affirmative sensitive responses and several with negative sensitive responses. This pattern explains why the intervals in Figure 2.2 appear to be narrower for these studies. Meanwhile, the widths of the intervals in Figure 2.2 are governed by the maximal possible proportion of liars in Figure A.1 and Figure A.2, which in turn depends on the prevalence of different control and sensitive items that varies across studies.

One important implication is that the crude bound is more informative when the control items and the sensitive item are of the same type, which may give rise to positive correlation among different items. In a list experiment included in a large-scale administrative survey in Malesky, Gueorguiev, and Jensen, 2015, items on the control and treatment lists are different activities that firms engage in to expedite the steps needed to receive their investment license/registration certificate. Another notable example is Meng, Pan, and Yang, 2014, who study the factors that Chinese government officials consider when making local policy and expenditure decisions. Control items include considerations such as local administrative expenditures, influence in attracting foreign investment and scope of the migrant population, and the sensitive items are suggestions from residents expressed through formal state institutional channels and those through the Internet. Given the items included in these list experiments, it's not surprising that the crude bounds are informative, as is clear from Figure 2.2.

Simulated List Experiments

I now investigate the determinants of the length of the intervals through several simulated list experiments. The benefit of looking at these artificially generated list experiments is that I have full knowledge of the data generating processes, which is not the case for list experiments in published studies. I focus on the differences in the data generating processes and the resulting differences in the interval lengths, and not absolute lengths.

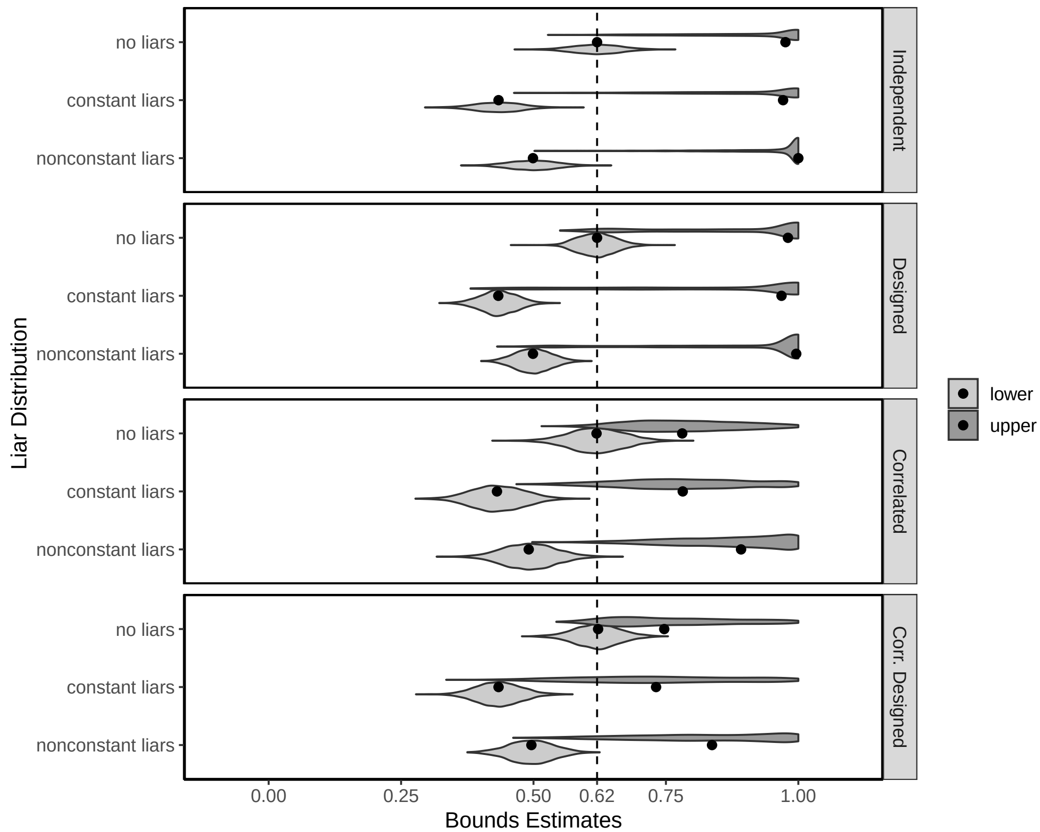
I consider a sensitive item of high prevalence (0.62) and list experiments of length 4 with 2000 respondents in total, and vary the design of the control-item list and the distribution of liars conditional on the number of control items answered affirmatively.

I study four list structures. In an "independent" list, all control items are independent with prevalence $(1/6, 1/2, 2/3, 2/3)$, and independent from the sensitive item. In a "designed" list, there are high- and low-prevalence control items (0.85 and 0.15), as well as two negatively correlated control items, with prevalence 0.5 and correlation -0.6. These parameter choices are borrowed from Ahlquist, 2017. I modify the "independent" lists and "designed" lists by adding 0.1 to the correlation of each pair of (control and sensitive) items, and refer to them as "correlated" and "correlated design" lists.

I look at three different liar distributions. In the case of "no liars", the lower bound—the difference-in-means estimate—consistently estimates the prevalence of the sensitive item. In the case of "constant liars", I assume 30% of the respondents who favor the sensitive item answer the question as if they do not favor it. Lastly, in the case of "nonconstant liars", I assume $30\% \times 0.8^k$ liars for respondents who answer affirmatively to all but k control items.

Figure 2.3 displays the densities and the medians of the crude bound estimates (lower bounds in lighter grey and upper bounds in darker grey) for 2000 simulated datasets for each of the 12 types of list experiments under consideration. As expected, in the case of "no liars", the lower bounds consistently estimates the prevalence of the sensitive item. In the case of "constant liars" and "nonconstant liars", the lower bounds underestimate the prevalence of the sensitive item by the proportions of liars in the population, whereas the interval estimates cover the prevalence. If the sensitive item is truly independent of control items, then the prevalence of sensitive item conditional on the number of control items answered affirmatively is flat. And as the top half of the figure shows, I have trivial upper bounds most of the time for independent and designed lists. On the other hand, if there is positive correlation between the sensitive item and control items, then the prevalence of sensitive item conditional on the number of control items answered affirmatively is increasing. And as the bottom half of the figure shows, I have non-trivial upper bounds for

Figure 2.3: Bound Estimates for Simulated List Experiments



Note: The figure presents the densities of lower bound and upper bound estimates for 2000 simulated datasets for each of the 12 types of list experiments. Each shaded area has measure one. Lower bound densities are shown in lighter grey and upper bound densities are shown in darker grey. Black dots are the median lower bound and upper bound estimates.

correlated and correlated design lists.

I present further simulations in the supplementary materials. The widths of the bounds are decreasing in the correlation between list items (Figure A.3 and A.4), consistent with our observation in Section 2.2 that decreasing the maximal liar ratio while keeping the prevalence fixed narrows the bounds. On the other hand, whether the sensitive response is affirmative or negative has no effect on the bound widths (Figure A.6), suggesting the difference between studies with affirmative and negative sensitive responses in Section 2.3 are due to differences in other aspects of these list experiments. Finally, list experiments with lower-prevalence items tend to produce slightly wider bounds in my simulations (Figure A.5).

Summary

By construction, the bounds will be nontrivial if and only if the proportion of truth-telling supporters is higher for respondents answering affirmatively to all J control items than $k < J$ control items for some k. The stronger the tendency that the proportion of truth-telling supporters increases in the number of control items answered affirmatively (especially when it approaches the boundary), the more informative the crude bounds are, as applications illustrate.

To put in primitive terms, the simulation exercise shows that positive correlation between the sensitive item and control items overall leads to narrower bounds, which is consistent with my previous observation that the width of the crude bound is narrower when items are of the same type. Absent such positive correlation, I expect uninformative bounds. For a list experiment with auxiliary question asking about the sensitive item directly, in addition to visualization in the form of Figure 2.1, researchers can also learn about the correlation by examining the correlation between control respondents' answers to list experiment and their answers to the auxiliary question.[13]

2.4 Discussion

Strengthening the Relaxed Liars Assumption

The relaxed liars assumption is a weaker assumption that the no liars assumption, leaving open the possibility that assumptions lying in between may give a tighter bound while accommodating the potential presence of liars. One such assumption states that the proportion of liars is increasing in the number of control items answered affirmatively:

$$\frac{p_{kL}}{p_{kL} + p_{kT}} \leq \frac{p_{mL}}{p_{mL} + p_{mT}}, \ \forall 0 \leq k < m \leq J - 1. \tag{2.6}$$

Without further information or assumption about the proportion of non-supporters, p_{mL}, one can weaken the inequality by replacing $p_{mL}/(p_{mL} + p_{mT})$ with $(p_{mL} + p_{mN})/(p_{mL} + p_{mN} + p_{mT})$. Hence, the improvement of this stronger assumption over the relaxed liars assumption depends on the relationship between $(p_{mL} + p_{mN})/(p_{mL} + p_{mN} + p_{mT})$ and $(p_{JL} + p_{JN})/(p_{JL} + p_{JN} + p_{JT})$. Since $(p_{JL} + p_{JN})/(p_{JL} + p_{JN} + p_{JT})$—one minus the proportion of truth-telling supporters conditional on all control items answered affirmatively—is smaller for many list experiments in Section 2.3, the gain is limited for these applications.

[13] This requires the propensity to lie about the sensitive item under the auxiliary question to be uncorrelated with the number of control items answered affirmatively under the list experiment for control respondents.

Nonstrategic Respondent Error

By looking at violations of the no liars assumption, I essentially restrict attention to strategic misreporting. In addition to strategic misreporting, an inattentive respondent engaging in survey satisficing may choose randomly an answer, the first answer or the last answer from all options available, suggested by evidence from placebo list experiments (Holbrook and Jon A. Krosnick, 2010; Ahlquist, Mayer, and Jackman, 2014; C. P. Kiewiet de Jonge and Nickerson, 2014; Ostwald and Riambau, 2018).[14] R Michael Alvarez et al., 2019 find in two datasets with screener questions, difference-in-means estimates for respondents failing screener questions are significantly different from those for attentive respondents.[15]

Nonstrategic respondent error can lead to violations of no design effect, no liars, and my relaxed liars assumption, result in bias to the difference-in-means estimator and maximum likelihood estimator à la Imai, 2011 (Ahlquist, 2017), and undermine the validity of the bounds developed in this paper. For online surveys, I recommend that applied researchers use screener questions—instructional manipulation checks (Oppenheimer, Meyvis, and Davidenko, 2009; Berinsky, Margolis, and Sances, 2014) or conventional instructed-response items (R Michael Alvarez et al., 2019)—to identify inattentive respondents so as to minimize such biases whenever possible. For face-to-face or telephone surveys (where survey satisficing is arguably less serious), screener questions are infeasible and researchers may have to rely on response time or other indicators to identify inattentive respondents.

Heterogeneous Lying Patterns

Some subgroups of respondents may be more truthful than others. If a respondent characteristic is correlated with both the number of control items answered affirmatively and lying about the sensitive item, then depending on the lying pattern and the magnitude of such correlation, it could potentially violate the relaxed liars assumption. For example, suppose respondents with less education are less likely to realize when ceiling effects are occurring and are more truthful as a result. Then the relaxed liars assumption will be violated if they are disproportionately more likely to answer affirmatively to all control items, and the magnitude is large enough

[14] A placebo list experiment replaces the sensitive item in a standard list experiment with an item whose prevalence is either known or estimable.

[15] In speculation to why such inflation/deflation might happen, C. P. Kiewiet de Jonge and Nickerson, 2014 mentions "[g]iven that the ICT tends to require greater cognitive effort than direct questions (particularly in the case of live-interviewer modes) and some respondents engage in satisficing strategies." Ahlquist, Mayer, and Jackman, 2014 find some suspicious respondents display straight-lining patterns.

to offset the strategic incentive underlying the assumption. Researchers can check whether respondents answering affirmatively to all control items are very different in composition from other respondents, and make changes to the control items if possible (for example, after a pilot study). If there is still concern, researchers can apply the method to each demographic group separately (estimation precision will be lower).

2.5 Conclusion

Standard analysis of list experiments requires a strong assumption on respondents' truthfulness toward sensitive issues, known as no liars. In this paper, I derive bounds for the prevalence of sensitive behaviors or attitudes under a weaker behavioral assumption and apply my method to an example on anti-immigration attitudes and a broad set of list experiments in the literature. The widths of the bounds vary substantially across studies. In some cases, especially when items on the list are of the same category, I can still reach substantive conclusions without putting full faith on the no liars assumption. In other cases, the full power of the no liars assumption is needed to pin down the prevalence of the sensitive behaviors or attitudes. I suggest that applied researchers compute the bounds as a guard against potential existence of liars in list experiments.

List experiments are designed to incentivize more truth telling from respondents than direct questions, but how successfully any particular application achieves it is not clear.[16] In particular, a small incremental support for the sensitive item over direct questions can mean either a small social desirability bias, or an unsuccessful inducement of truth telling. The method proposed in this paper is a first attempt to accommodate the presence of liars. The bounds exploit underlying strategic incentives and information in list experiments, but can be uninformative depending on applications. When there are resources, applied researchers can also employ other strategies such as the randomized response technique and compare estimates to gain confidence.

The method I propose is an addition to the list experiment toolbox and applicable to a large number of list experiments. But post-data-collection methods should not substitute or be prioritized over efforts to design good list experiments in the first place. Applied researchers should carefully construct the list items so that the question looks natural and the sensitive item stands out to a lesser extent, and follow

[16] Another issue worth pointing out is that there is potentially a file drawer problem in list experiment research, like many other areas of scientific research.

other recommended practices discussed in Blair and Imai (2012) and Glynn (2013). Meanwhile, when the list features a positive correlation between the sensitive item and control items overall, researchers can exploit this opportunity to estimate bounds robust to the presence of liars.

Chapter 3

PAYING ATTENTION TO INATTENTIVE SURVEY RESPONDENTS

Alvarez, R. Michael et al. (2019). "Paying Attention to Inattentive Survey Respondents". en. In: *Political Analysis* 27.2, pp. 145–162. DOI: 10.1017/pan.2018.57. URL: https://www.cambridge.org/core/journals/political-analysis/article/paying-attention-to-inattentive-survey-respondents/BEDA4CF3245489645859E7E6B022E75A.

3.1 Introduction

Over the last two decades the environment in which respondents participate in surveys and polls has changed, with shifts from interviewer-driven to respondent-driven surveying, and from probability to nonprobability sampling. It is still not known precisely how these changes in the survey environment have affected the quality of survey response. Also, response rates for traditional polling have been declining dramatically (Atkeson, Adams, and R. Michael Alvarez, 2014). These changes have focused the attention of survey methodologists on data quality, and on the motivation and engagement of survey respondents. These questions are important in political science where surveys are a primary tool for testing theories of political behavior, and where many researchers use new methodologies like opt-in non-probability samples with national coverage (e.g. Cooperative Congressional Election Study, CCES), as well as survey respondent workforces such as mTurk or Google Consumer Surveys (for discussions of these survey techniques see Berinsky, Huber, and Lenz 2012; Ansolabehere and Brian F Schaffner 2018).

Online panels, and other new technologies such as Interactive Voice Response (IVR), Computer Assisted Personal Interviewing (CAPI), and Address Based Sampling (ABS), have been behind the change from predominantly interviewer-driven survey environments, face-to-face (FTF) and telephone, to respondent-driven ones (Atkeson and Adams, 2018; Dillman, Smyth, and Christian, 2009)). The presence of an interviewer enforces some control over the pace of the interview and social dynamics are believed to increase respondent engagement, while the lack of one gives control of the survey to the respondent, which allows for more reduced engagement.

One consequence of these technological changes is that survey respondents in these environments may be less attentive to survey questions. Some respondents may pay little attention to the questions or their responses, while others may deliberately misrepresent their behavior or preferences (Atkeson and Adams, 2018). This is a cause for concern, as well-considered responses are necessary for quality survey data. The expectation in a survey environment is that the respondent is mindful in the survey process—reading or listening, and then engaging in cognitively to provide a meaningful answer to every survey question. Lack of attentiveness may be a source of nonsampling bias and response error, and a contributor to total survey error (Groves and Lyberg, 2010). This may increase the amount of noise in the data, producing inaccurate estimates, and hampering our ability to test hypotheses with precision and accuracy.

Alternatively, noisy data may be inherent in survey research because it may reflect the ambiguity, disinterest, inattentiveness, and distraction that pervades citizen interest in politics and policy (R. Michael Alvarez and Brehm, 2002). In this way, including inattentive respondents in surveys may be important because citizen non-attitudes are prevalent in the public on any particular issue or topic, and accounting for non-attitudes might be crucial for making accurate inferences about research questions. Therefore, it is important to study how to identify engaged and disengaged respondents, and the implications of their responses for testing theories of political behavior.

In many circumstances, simple direct questions may not adequately elicit useful information from respondents. This is particularly true for sensitive issues, where eliciting truthful answers directly is not feasible due to social desirability bias (E. E. Maccoby, N. Maccoby, and Lindzey, 1954; Edwards, 1957; Fisher, 1993). Researchers have developed indirect approaches, such as the randomized response technique (Warner, 1965) and the list experiment (J. D. Miller, 1984), for measuring sensitive behavior and attitudes via opinion surveys (for a recent review see Rosenfeld, Imai, and Shapiro 2016). These techniques involve indirect questions that have longer question wording, more complex structure, and are more cognitively demanding. These types of questions have not been investigated in the past in relation to respondent attentiveness but, given the complexity of popular approaches for measuring sensitive behavior, we expect inattentive respondents to provide much less accurate and consistent information on their sensitive dispositions than those who carefully consider survey questions.

One strategy used in online surveys to detect inattentive respondents is the inclusion of attention checks (i.e. screeners for attention, also called trap or red herring questions), such as instructed-response items and instructional manipulation checks (IMCs), which instruct respondents to answer a question in a specific way.[1] Oppenheimer, Meyvis, and Davidenko, 2009 demonstrated that attention checks in the form of IMCs are effective at detecting participants who are not following instructions, increasing statistical power and data reliability. Berinsky, Margolis, and Sances, 2014 documented that numerous studies used IMCs from 2006 to 2013, and recommended using multiple IMCs to measure attention. In some research designs, these techniques are used as filters, with failing respondents eliminated from the survey, while in others the information is used to assess data quality.[2] In our study, we use instructed-response items as attention checks to identify inattentive respondents. We examine the characteristics of respondents who failed our trap questions and compare them to respondents who did not. We then examine how inattentive and attentive survey respondents answered a series of questions about political attitudes and behavior, including a double list experiment—-an indirect questioning technique introduced in Droitcour et al., 1991, which combines two standard list experiments on the same sensitive issue to improve efficiency and gain additional diagnostic opportunities (Glynn, 2013).

3.2 Survey Satisficing and Attention Checks

Although survey researchers want their respondents to be engaged in the survey process, it is likely that some respondents may not be completely engaged. When faced with demanding information-processing tasks some respondents will expend only the minimum amount of effort to provide a response. In psychology Simon, 1956 described this as satisficing. In the context of the survey response process, respondents who satisfice may not search their memory completely or may not fully understand the question, and in general they will take a superficial approach to the question-answer format (Jon A Krosnick, 1991). In extreme cases, respondents may not even pay attention to a question, and engage in random guessing (Oppenheimer, Meyvis, and Davidenko, 2009; Jones, House, and Gao, 2015; Berinsky, Margolis, and Sances, 2014).

A strategy for identifying inattentive survey respondents involves embedding attention checks in carefully selected locations in the survey instrument. One set

[1] We use the terms "attention checks" and "trap questions" interchangeably throughout the paper.

[2] Eighty percent of papers documented in Berinsky, Margolis, and Sances, 2014 exclude failures from their analyses.

of attention checks are instructed-response items that ask respondents to provide a specific response and are often part of a grid question or a group of questions with the same scale (e.g., rows in a grid instructing respondents to "please select 'strongly disagree' for data quality control"). Instructed-response items evaluate respondents' compliance with simple and concise instructions and have been used in several previous studies as attention indicators (L. K. Barber, Barnes, and Carlson, 2013; Ward and Pond, 2015; Bowling et al., 2016). Some inattentive respondents may pass the attention check by chance, but this problem can be mitigated with the inclusion of multiple such items in the survey. A closely related technique involves infrequency scales or bogus items. These are items on which all or virtually all attentive respondents should provide the same response, e.g. "I have never used a computer" (Huang et al., 2012). In their research, Meade and S. B. Craig, 2012 conclude "we strongly endorse bogus items—or, preferably, instructed response items (e.g. "Respond with 'strongly agree' for this item")—for longer survey measures."

Another popular type of attention check is the instructional manipulation check (IMC) first introduced by Oppenheimer, Meyvis, and Davidenko, 2009. A typical IMC evaluates whether respondents read and follow instructions within a lengthy question prompt—instructions taking precedence over other requests for information made elsewhere in the question's text (e.g. may include phrases such as "please ignore the rest of the question and select options A and C"). In contrast to instructed-response items, which would not make sense if respondents ignore the part of the question requesting specific responses, IMCs are perfectly valid questions without the phrase containing the instruction, and could therefore be more easily misunderstood by inattentive respondents. Both type of attention checks may be used for general inattention detection, but IMCs are particularly useful for survey experiments where "the manipulation of a study is hidden in similar text somewhere else in that study" (Curran, 2016) and instructed-response items are particularly useful for grid and check-all-that-apply questions.

Besides attention checks, previous research suggests that indices capturing response patterns, such as overall response duration and frequency of non-attitudes, may provide valuable information about respondent attentiveness (Johnson, 2005; Huang et al., 2012; Meade and S. B. Craig, 2012; Maniaci and Rogge, 2014; Curran, 2016). We consider consistency of expressed preferences, propensity to select the same response to multiple contiguous questions (i.e. straightlining behavior), and response time as additional validation for attention measured based on multiple instructed-response items.

Depending on the reason for the failure to pass attention checks, trapping and removing inattentive respondents from the sample may or may not be a reasonable approach for dealing with respondent satisficing (Berinsky, Margolis, and Sances, 2014; Downes-Le Guin, 2005; Zagorsky and Rhoton, 2008). If the motives behind satisficing behavior correlate with respondent characteristics influencing the outcome of interest, then listwise deletion of inattentives could lead to inaccurate inferences as inattentive responses would not be missing completely at random (MCAR; Little 1992). Past research suggests that respondents who fail trap questions are younger (Kapelner and Chandler, 2010; Berinsky, Margolis, and Sances, 2014), more likely to male (Kapelner and Chandler, 2010; Berinsky, Margolis, and Sances, 2014), less educated (Kapelner and Chandler, 2010), and more likely to be non-white (Berinsky, Margolis, and Sances 2014, but see Anduiza and Galais 2016). These findings are partially consistent with our results and suggest that caution should be exercised in deciding whether to keep or drop inattentives. If unobserved factors influence both attentiveness and the mechanism behind the outcome of interest, then it would also be inappropriate to treat inattentive responses as missing at random (MAR) conditional on covariates. Missingness resulting from the process of removing inattentives would then be missing not at random (MNAR) and therefore nonignorable, and it may not be possible to successfully address it via conventional model-based imputation procedures (King et al., 2001; Pepinsky, 2018).

Ultimately, as we discuss toward the end of the paper, how to deal with inattentiveness comes down to a comparison of costs and benefits associated with keeping inattentive respondents in the sample. Designing a survey where attention checks are used, and both attentive and inattentive survey respondents complete the survey, produces noisy data at significant cost. In that context, statistical adjustments may become necessary to account for differences in attentiveness between respondents. On the other hand, designing a survey where inattentive survey respondents are dropped from the sample lowers survey costs considerably (few respondents need to be interviewed for the entire survey), reduces noise, but risks producing an unrepresentative sample that will require post-hoc statistical adjustment to produce population-level inferences.

3.3 Data and Methods

Data

We use data from an online survey of 2,725 California adults conducted in July 18-30, 2014 using a sample recruited by Qualtrics through the e-Rewards panel.[3] Recruitment into the e-Rewards online panel was carried out using a double opt-in process, whereby "[a]fter receiving a personalized email invitation to join the e-Rewards program, individuals must opt-in and agree to provide truthful and well-considered answers [. . .]. After the first opt-in during the enrollment process, the individual is sent a follow-up e-mail confirmation that requests for him/her to click on a link to validate opt-in. [. . .] Once a member has completed the double opt-in process, they are then eligible to begin receiving survey invitations" (e-Rewards 2008, p. 3).

In this survey panel participants were invited via email to complete a 20-minute respondent-driven online questionnaire, which was designed and implemented using Qualtrics survey software.[4] Respondents could choose between English and Spanish versions of the questionnaire.[5] The data collection process began on July 18, 2014 and concluded on July 30, 2014 when the target sample size of 1,700 complete and attentive responses was reached. Individuals that failed to meet gender, age, and education quotas were filtered out at the beginning of the survey after completing a brief screener section assessing basic demographics.[6] Respondents who reached the end of the survey answered additional demographic questions that were later used to construct survey weights. Those weights are not used in this paper as the demographic data used in calculating weights are only available for respondents who completed the entire survey—i.e. are not available for respondents that failed trap questions. Estimates reported in this paper apply to the sample at hand, and do not necessarily reflect the California adult population.

[3] The sample of 2,725 respondents includes 1,750 responses from individuals who completed the entire questionnaire and 975 partial responses from individuals who were filtered out before the end of the survey after failing a trap question. Data and code for replicating the analyses reported in this paper are available in Dataverse.

[4] Figure B1 in Appendix B shows a screenshot of a sample email invite.

[5] Only 21 respondents (less than 1%) chose to see the survey in Spanish.

[6] Survey quotas were selected as to ensure a minimal number of complete responses within pre-defined gender, age, and education categories. To ensure the timely completion of the data collection process, we allowed for over-representation of age/education groups. Quotas were removed on July 29, 2014, after more than 1,600 complete responses had been collected, to speed up the conclusion of the data collection process.

Identification of Inattentive Respondents

Research on attention checks suggests that many respondents within any research design may be inattentive and hence be more likely to satisfice. Estimates suggest the range of failure for attention checks is quite large, from as low as 8% to 50% of respondents (J. Miller, 2006).

In our study, we had three instructed-response items that appeared at different points in the survey and used different means by which to assess whether a subject was paying close attention to the survey questions. Figures B2-B4 in Appendix B show screenshots of desktop versions of these questions. The three trap questions (TQ) were:

- TQ 1: Located immediately below a check-all-that-apply question on participation in twelve political activities. It instructed respondents to type the word "government" inside a text box. Answers were coded as correct if they contained the term "gov."
- TQ 2: Grid question that asked respondents to report support or opposition to several policies. In one of the rows, respondents were instructed to select "I'm indifferent" for quality purposes. The location in the grid of the row containing the instructed-response item was randomly determined.
- TQ 3: Grid question that asked respondents to agree/disagree with statements indicating varying levels of tolerance toward other people's views. In one of the rows, respondents were instructed to select "Disagree Strongly" for quality purposes. The location in the grid of the row containing the instructed-response item was kept fixed at the bottom.

Respondents that failed each instructed-response item were filtered out. Nonetheless, incomplete responses provided by respondents who eventually failed a trap question were recorded up to the moment they were dropped from the survey. This allowed us to use responses to questions preceding trap questions to compare the attributes of those who did and did not survive each attention check, and subsequently evaluate how inattentiveness distorts the observed distribution of political attitudes and behavior. While inattentive respondents may have passed some of these attention checks by chance—particularly the ones embedded within a grid—the chance of surviving all attention checks should be considerably higher for attentive respondents than for those engaging in satisficing behavior. In the rest of this article, we refer to respondents that passed all checks as attentives, and to the rest as inattentives.

In the next section, we provide some basic information that profiles the characteristics of attentive and inattentive survey respondents. We then validate our attentive–inattentive distinction by considering alternative measures of response quality (response time, item nonresponse, and response consistency) for attentives and inattentives. We then provide evidence on systematic differences between attentive and inattentive respondents in terms of their reported political attitudes and behavior.

3.4 Results

In total, slightly more than one third (36%) of respondents were inattentive, as they failed to pass one of the trap questions (see Table B1 in the supplementary materials).[7] The large incidence of failure to pass attention checks suggests that inattentiveness is a common phenomenon in respondent-driven surveys like this one.

Among respondents that failed the first attention check, the most common behavior was leaving the text box empty and skipping the question (displayed by 96% of those that failed TQ 1). Among respondents that failed the second attention check (which instructed individuals to select "I'm indifferent" along a scale that also included the options "I don't know", "Oppose," and "Support"), the most common behavior was selecting "Support" (selected by 44% of those that failed TQ 2, and coinciding with the modal response to all other items included in the grid). We did not find that the order of the instructed-response item in the grid, which was randomized for the second attention check, had a significant influence on failure rates. It is possible that inattentives passed TQ 2, as respondents were instructed to report a non-attitude ("I'm indifferent"). This was not the case for TQ 3, where the instructed response represented a definite stance ("Disagree Strongly") and was an uncommon answer to other items in the grid. Among respondents that failed the third attention check, 60% selected "Neither Agree nor Disagree," a non-attitude, in response to TQ 3.

In addition to being commonplace, inattentiveness does not occur completely at random. Those who passed all attention checks differ demographically from those that failed (see Table B2 in the supplementary materials). Less educated and younger respondents are more likely to fail, consistent with Berinsky, Margolis, and Sances,

[7]The breakdown of the failure rates over the three trap questions was as follows: 21%, 8%, and 6% of respondents were filtered out after failing to pass the first, second, and third attention checks, respectively.

2014 and Kapelner and Chandler, 2010, but we find no gender differences in attentiveness, which contrasts with previous studies. We also find that inattentives interact with the survey in specific ways: they answer questions more quickly, consistent with Oppenheimer, Meyvis, and Davidenko, 2009; are more likely to answer "don't know" and are more likely to straightline (see Table B3 in the supplementary materials). These latter two results are consistent with past work on satisficing by survey mode (see Atkeson and Adams 2018).

Are the three trap questions similar in terms of their ability to detect inattentives? Since respondents that failed an attention check were filtered out, we are not able to compare the characteristics and behavior of respondents who fail each trap question but pass the rest. We can, nonetheless, learn much about the operation of each trap question by evaluating whether two reasonable expectations hold within the following nested groups of respondents: those that see TQ 1 (the entire sample), those that see TQ 2 (including only respondents that pass TQ 1), and those that see TQ 3 (including only respondents that pass TQ 2).

Within each of these groups, we examine whether respondents that pass the next trap question look and behave differently than those that fail. Because of differences in the composition of these three groups, however, differences between those that fail and pass each check cannot be compared between groups. But given that individuals who pass more checks are likely to be more attentive than individuals who pass fewer checks, we expect that a typical respondent surviving all checks, for instance, will not be less attentive than the typical respondent in the entire sample. Conditional on this testable assumption (which can be validated based on auxiliary information on attentiveness collected for all respondents before exposure to TQ 1, such as response speed up to that point), if the three trap questions operate similarly, then observed differences between respondents that pass and fail each check should become progressively smaller as we move from considering all respondents to only considering those that get to see TQ 3, as the most inattentive respondents will have been filtered out before reaching the last attention check. Observed differences between respondents who pass and fail TQ 1, for instance, should be more pronounced than differences between respondents who pass and fail TQ 3, as respondents that get to see TQ3 should display higher (and more homogeneous) levels of attentiveness, on average, than the broader group of respondents including individuals that eventually failed TQ 1 or TQ 2.

Using the information provided in Tables 2 and 3, we find that the younger and less educated respondents are, the more likely they are to fail TQ 2 and 3. Respondents who fail these two attention checks, in turn, spend less time considering survey questions, report higher rates of non-attitudes, and are more likely to display incomplete and/or intransitive preferences over policy options. These patterns, however, are less pronounced in the case of TQ 1. Differences between respondents that pass and fail TQ 1—in terms of characteristics listed in Table B2 and behaviors listed in Table B3—are no larger than differences between respondents that pass and fail TQ 2 and 3, respectively, a finding that contradicts our second expectation.

These results indicate that TQ 1 (an open-ended instructed-response item that instructed respondents to type the word "government" in a text box, located immediately below a check-all-that-apply question with numerous response alternatives) operates differently than TQs 2 and 3 (instructions given in rows of two separate grid questions, which instructed respondents to select specific responses along a labeled scale). While we cannot evaluate why different types of trap questions filter out different respondents, we argue that TQ 1 filtered out respondents for reasons having little to do with inattentiveness, which would explain why respondents that fail TQ 1 differ little in terms of demographic attributes and earlier interaction with the survey instrument relative to those that pass TQ 1. Unlike instructions given in TQ 2 and 3, instructions given in TQ 1 did not state that the request made in the question was for quality purposes. It is possible that many respondents found TQ 1 senseless and as a result decided to ignore the request. Another possibility is that some respondents may have failed to notice the text-box question, as it was located immediately below a check-all-that-apply question with numerous response alternatives (see Figure A2 in Appendix A), rather than in a stand-alone page (as most other questions in the survey). Both of these explanations help account for the observation made before that most respondents that failed TQ 1 did so not because they wrote something other than "government" in the text box, but because they left the text box empty.

In sum, we find that respondents that pass all attention checks differ from those that fail, but these differences are more pronounced for closed-ended grid-style instructed-response items than for the open-ended text-box-style instructed-response item. This suggests that some trap questions may be more reliable than others and could make a big difference in terms of who gets filtered out or flagged as inattentive, as could slight differences in question wording (such as whether the question is designated with the intent of verifying response quality or whether it is displayed in the same page as other survey questions). We leave further exploration of this question for future research.

Direct Questioning Techniques

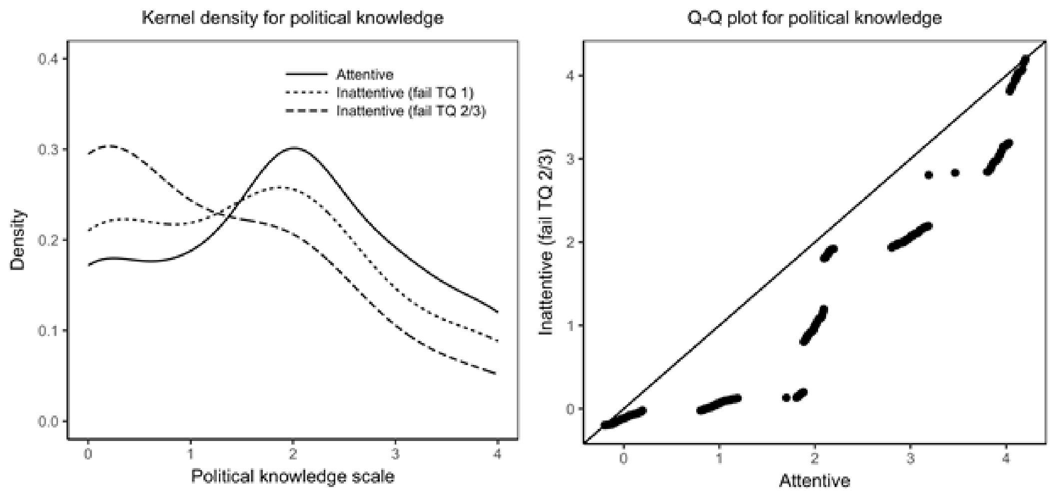
To consider the extent to which attentives and inattentives provide different answers to direct questions, we examine their responses to closed-ended questions about their political knowledge, self-reported political activity, and scale placement on policy issues. These measures may demonstrate whether political interest is associated with attentiveness.

We used responses to four questions about factual knowledge of California politics to construct a 0–4 political knowledge scale. The first two knowledge questions asked about the majority party in the State Senate and Assembly. The other two asked about majorities required for passing constitutional amendments and for raising taxes. Figure 3.1 shows the distribution of the political knowledge scale for attentive and inattentive respondents. These results suggest that inattentives miss factual knowledge questions more frequently. In the case of the two questions on majority requirements, inattentives are about as likely as attentives to select “I don’t know” (34% and 37% of the time for each question, respectively, the same as attentives). More inattentives fail these knowledge questions because they are more likely to select the wrong percentage (in particular, 6% of inattentives report that unanimity is required for each decision to pass, compared to only about 2% of attentives). In the case of the knowledge items asking about majority party in the State Senate and Assembly, inattentives are both considerably more likely to report “I don’t know” and to get the answer wrong by selecting “Republican.”[8]

We then used responses to a check-all-that-apply question about participation in twelve political activities to construct a 0-12 political participation scale that was asked immediately before exposure to the first attention check (for a screenshot, see Figure A2 in the Appendix). Listed activities included voting in national and statewide elections, other conventional forms of involvement, and involvement in unconventional acts. Figure 3.2 shows the distribution of the political participation scale for attentive and inattentive respondents. Summary statistics of self-reported participation in each activity are indicative of non-random selection of a small num-

[8] In the question about the majority party in the State Senate, 29% of inattentives report “I don’t know” and 21% incorrectly select “Republican,” compared to 20% and 14% of attentives, respectively. In the question about the majority party in the State Assembly, 35% of inattentives report “I don’t know” and 20% incorrectly select “Republican,” compared to 29% and 10% of attentives, respectively.

Figure 3.1: Attentiveness and political knowledge

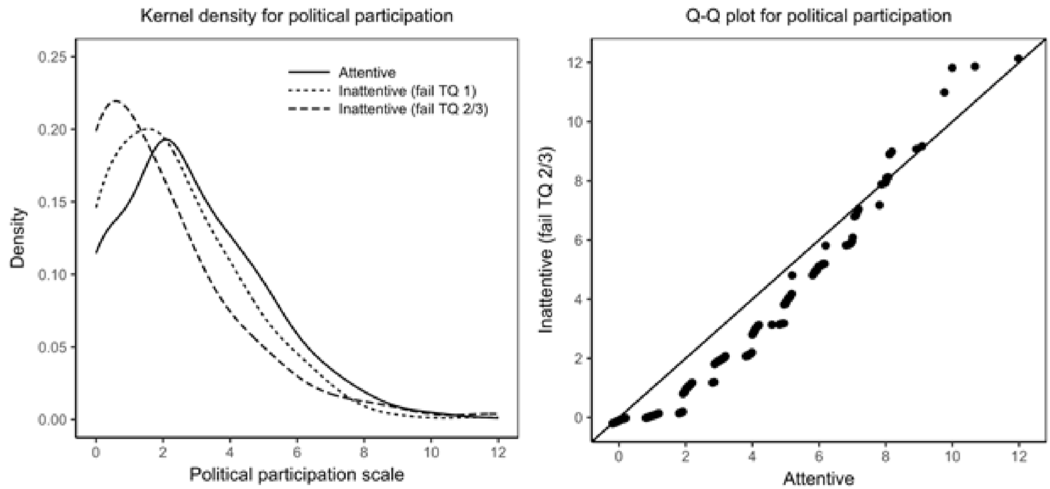


Note: The plot on the left displays the kernel distribution of political knowledge (measured on a 0-4 scale) among respondents that passed all attention checks (solid line), failed the first check (dotted line), and failed the second or third check (dashed line). The quantile-quantile plot on the right compares quantiles of the distribution of political knowledge between respondents that passed all checks (horizontal axis) and those that failed the second or third check (vertical axis).

ber of response alternatives by inattentives, rather than entirely haphazard choices, as these respondents were consistently more likely to select common conventional activities (e.g. voting and signing petitions), than more demanding activities (e.g. working for campaigns, attending political meetings, and donating) or unconventional ones (e.g. participation in protests and sit-ins). These results suggest that inattentives report participating in fewer political activities due to lower levels of political engagement compared to attentives.

Lastly, we used responses to six questions on support for liberal policies to construct a 13-point ideology scale (ranging from -6 to 6). Respondents were asked about support for the Affordable Care Act, repealing "Don't Ask, Don't Tell," providing a path to legal status and citizenship for undocumented immigrants, implementing stricter carbon emission limits, restricting the sale of semi-automatic and automatic weapons, and limiting NSA's collection of domestic phone records. We coded support for each policy as 1 for respondents selecting "support," -1 for those selecting "oppose," and 0 for those selecting "I'm indifferent" or "I don't know." Figure 3.3 shows the distribution of the ideology scale, constructed by adding up support across the six policies, for attentive and inattentive respondents. These results indicate that

Figure 3.2: Attentiveness and political participation

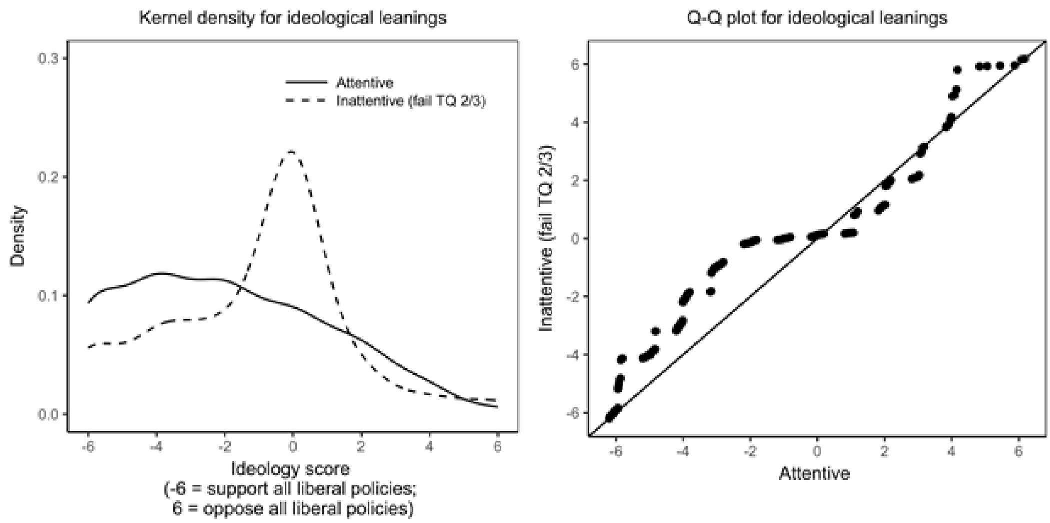


Note: The plot on the left displays the kernel distribution of political participation (measured on a 0-12 scale) among respondents that passed all attention checks (solid line), failed the first check (dotted line), and failed the second or third check (dashed line). The quantile-quantile plot on the right compares quantiles of the distribution of political participation between respondents that passed all checks (horizontal axis) and those that failed the second or third check (vertical axis).

inattentives report less conclusive stances toward policy issues. When looking at responses issue-by-issue, it is evident that inattentives do not select "I don't know" or "I'm indifferent" with equal probability for each policy issue. Among inattentives that select a non-attitude, we find that they are about twice or more as likely to select "I'm indifferent" than "I don't know" (particularly for repealing "Don't Ask Don't Tell" and the gun control item, where they are close to three times as likely to select "I'm indifferent" instead of "I don't know"; and with the exception of the health care law, where they are about as likely to select "I'm indifferent" as "I don't know"). These item-by-item findings suggest that the lower-intensity answers reported by inattentives reflect greater uncertainty about policy positions.

We estimated a series of linear regression models to evaluate whether attitudinal and behavioral differences between attentives and inattentives persist after controlling for demographic backgrounds of respondents that pass and fail attention checks (see Tables B4-B6 in the supplementary materials). The results of these analyses are consistent with those presented before. When inattentives are asked closed-ended questions about political knowledge, civic engagement, and opinions on policy issues, their responses reveal less knowledge, lower involvement, and weaker policy

Figure 3.3: Attentiveness and ideological leanings



Note: The plot on the left displays the kernel distribution of ideological leanings (measured on a -6 to 6 scale) among respondents that passed all attention checks (solid line) and those that failed the second or third check (dashed line). The quantile-quantile plot on the right compares quantiles of the distribution of ideology between respondents that passed all checks (horizontal axis) and those that failed the second or third check (vertical axis).

stances than attentives. These results do not allow establishing whether inattentives are shirkers who check fewer activity boxes and provide hasty responses, whether inattentiveness springs from genuine lack of interest in politics, or whether a mixture of both mechanisms is at play. A consideration of responses to indirect questioning may help to clarify the implications of these results for understanding inattentives.

Indirect Questioning Techniques

We have looked at direct questions on respondents' political knowledge, participation, and ideology placement. We now evaluate respondents' attitudes toward immigration through a double list experiment embedded in the survey.[9] The experiment was designed to measure Californians' support for two state chapters of national anti-immigrant organizations. To preserve the anonymity of these organi-

[9] In a standard list experiment, respondents in the control group see a list of control items and respondents in the treatment group see a similar list that also includes a sensitive item. A double list experiment consists of two lists presented to respondents, with different control items but the same sensitive item for respondents seeing the "treatment" version of each list. Respondents are randomly assigned to treatment (i.e. seeing the sensitive item) in the first or second list. Thus, in contrast to a standard list experiment, where only a subset of respondents (those in the treatment group) is exposed to the sensitive item, all respondents in a double-list experiment see the sensitive item at some point (either in the first or second list), leading to potential efficiency gains.

zations, we refer to them as Organization X and Organization Y. Organization X was described in the double list experiment as "advocating for immigration reduction and measures against undocumented immigration," and organization Y as a "citizen border patrol group combating undocumented immigration."

The double list experiment comprised two questions containing "list A" and "list B." The first question (list A) exposed respondents to a list of different groups and organizations in randomized order, and asked them to specify "how many of these groups and organizations you broadly support."[10] This was followed by a second list of different organizations (list B). The two questions provided the name and a brief description of all listed organizations, and were located after the first attention check, but before the second one.

Depending on their treatment status in the double list experiment, respondents saw different versions of list A and B that included or excluded the name and description of one of the two anti-immigrant organizations. Respondents assigned to the "control A - treatment B" condition were exposed to the sensitive item in list B; and those assigned to the "treatment A - control B" condition were exposed to the sensitive item in list A. Sensitive items displayed under either of these two experimental conditions were randomly assigned to respondents. Items displayed to respondents and the number of respondents assigned to each combination of experimental conditions and sensitive items are shown in Tables B7 and B8 in the supplementary materials, respectively.

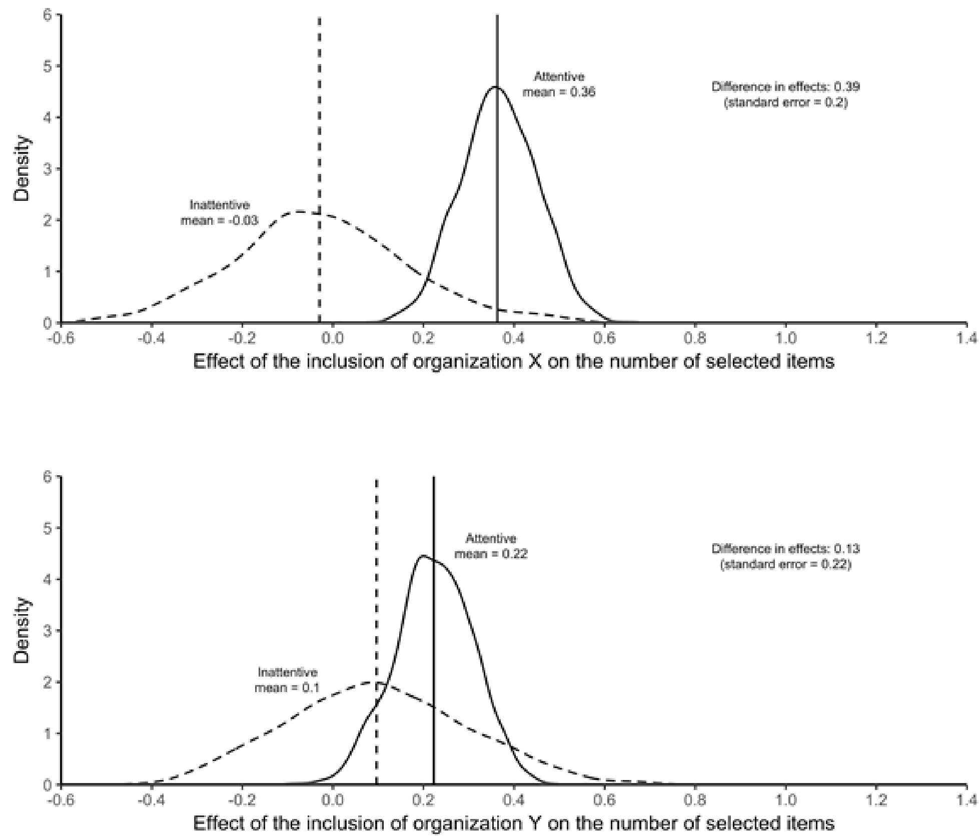
Possible responses to each list question comprised integers between zero and four under control and between zero and five under treatment (X or Y), representing the number of supported organizations. Under two assumptions—no design effect and no liars—the difference between the average response under treatment and control consistently estimates the level of support for the sensitive item (Imai, 2011; Blair and Imai, 2012).

Figure 3.4 provides a visualization of difference-in-means estimates for attentive and inattentive respondents, using list A responses. In the figure the vertical lines show the difference-in-means estimates in our sample and the curves show the distribution

[10] The exact instructions given to respondents were: "Below is a list with the names of different groups and organizations on it. After reading the entire list, we'd like you to tell us how many of these groups and organizations you broadly support, meaning that you generally agree with the principles and goals of the group or organization. Please don't tell us which ones you generally agree with; ONLY TELL US HOW MANY groups or organizations you broadly support. HOW MANY, if any, of these groups and organizations do you broadly support."

of these estimates in 1,000 bootstrapped samples. Attentives select 0.36 more items on average under the X-treatment (and 0.22 more items on average under the Y-treatment) than under control, whereas inattentives select a similar number of items under treatment and control. The distributions for inattentive respondents are more dispersed due to noisier responses and smaller sample sizes.[11]

Figure 3.4: Attentiveness and difference-in-means estimates (list A)



Note: The figure displays the distribution of difference-in-means estimates (mean response under treatment condition minus mean response under control condition) for attentive and inattentive respondents separately in 1,000 bootstrapped samples. Solid and dashed vertical lines show the difference-in-means estimates in the original sample for attentives and inattentives, respectively. The top-right figure in each plot reports the attentive-inattentive difference in difference-in-means estimates together with the corresponding bootstrapped standard error. Results for organization X and Y are shown on the top and bottom plots, respectively.

[11]This difference in dispersion is confirmed by a Levene's test (the F statistic from this test is 437 for organization X and 419 for organization Y).

Inattentive respondents on average choose fewer items under both control and treatment, with the difference more pronounced under treatment (see Table B9 in the supplementary materials). The decrease is consistent with two types of survey satisficing. First, not supporting a group can be an expression of a non-attitude. As we saw from previous sections, inattentive respondents are more likely to report attitudes in the middle of the scale as opposed to the extremes consistent with both shirking and disinterest. However, the list experiment suggests that shirking may be a primary factor. Alternatively, it could be that many inattentives did not pay attention to the list and chose a small number, especially the first option—zero in this case (see Table B10 in the supplementary materials for the distribution of responses).

This result suggests that inattention may account for artificial deflation due to list length documented in the literature, namely that estimators are biased due to the different list lengths provided to control and treatment (C. P. Kiewiet de Jonge and Nickerson 2014, 659).[12] C. P. Kiewiet de Jonge and Nickerson, 2014 find significant under estimation of the occurrence of a common behavior. In a recent paper, Eady, 2017 included a screener for a large-scale list experiment (n=24,020) and excluded respondents who failed the screener from the analysis. We reanalyzed data extracted from Eady's replication package (Eady, 2016) and found that respondents who failed the screener on average chose a smaller number of items under both control and treatment, with the difference more pronounced under treatment (see Table B12 in the supplementary materials). The difference-in-means estimate for those who passed the screener is 0.88, and the estimate for those who failed is 0.75 (difference: 0.13, s.e. = 0.05).[13]

In our double list experiment, results are dramatically different for list B, as shown in Figure 3.5. Attentives select 0.25 more items on average under the X-treatment (and 0.28 more items on average under the Y-treatment) than under control, which is very similar (and nearly identical) to what we showed in Figure 4 for the attentives. However, differences in means for the inattentives are now positive and large in magnitude (0.60 for organization X and 0.53 for organization Y).[14] The cross-list

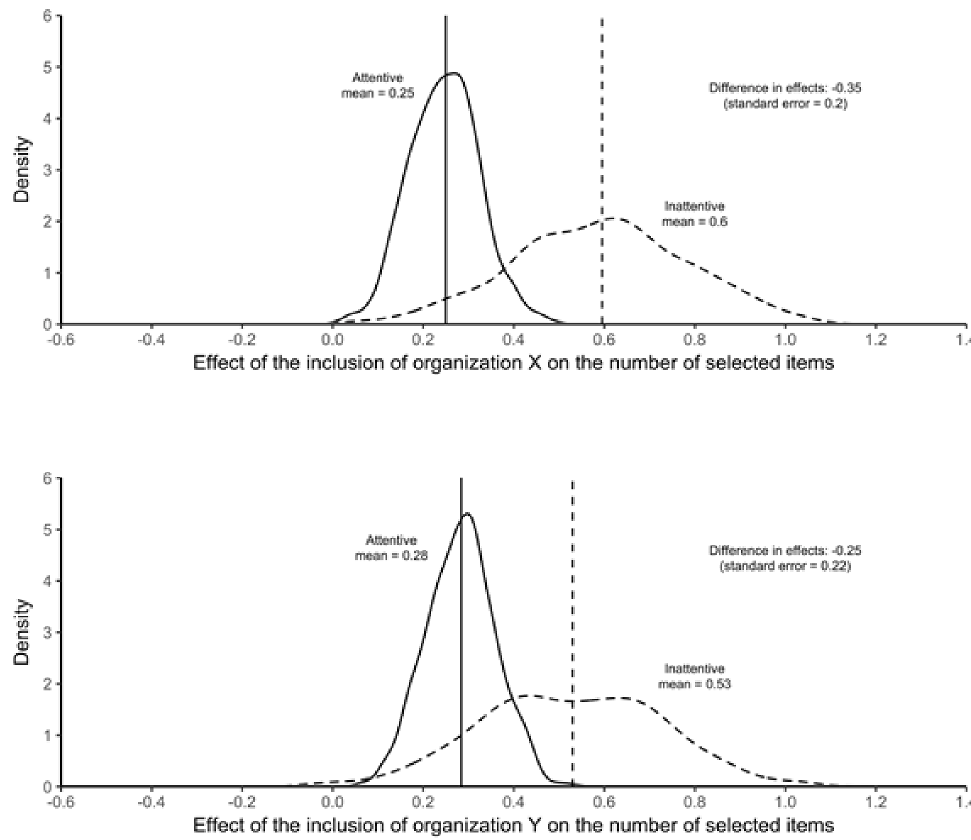
[12]Artificial deflation due to length effect can also arise if the inclusion of the sensitive item provides a strong contrast that reduces the attractiveness of control items on the list. We view the two causes as complementary and the magnitude attributable to each cause depends on survey environment as well as items on the list.

[13]In our list experiment, an affirmative latent response to the sensitive item is sensitive, whereas it is the opposite for Eady, 2017. This provides further evidence that the deflation is artificial and has nothing to do with social desirability bias.

[14]Again, the distributions for inattentive respondents are more dispersed, confirmed by a Levene's

difference for inattentives in terms of the difference-in-means estimates is 0.62 for organization X (s.e.= 0.34) and 0.43 for organization Y (s.e.= 0.35).

Figure 3.5: Attentiveness and difference-in-means estimates (list B)



Note: The figure displays the distribution of difference-in-means estimates (mean response under treatment condition minus mean response under control condition) for attentive and inattentive respondents separately in 1,000 bootstrapped samples. Solid and dashed vertical lines show the difference-in-means estimates in the original sample for attentives and inattentives respectively. The top-right figure in each plot reports the attentive-inattentive difference in difference-in-means estimates together with the corresponding bootstrapped standard error. Results for organization X and Y are shown on the top and bottom plots, respectively.

To see why the difference-in-means estimate for inattentive respondents is small for list A (Figure 3.4), but large for list B (Figure 3.5), notice the structure of the double list experiment is such that treated respondents for list A are now under control, and control respondents for list A now receive treatment. In the crosstabs shown in

test (the F statistic is 502 for organization X and 515 for organization Y).

Table B11 in the supplementary materials, we see that many respondents choosing 0 in list A continue to choose 0 in list B; also, those choosing the maximal number continue to choose the maximal number. This tendency is much more pronounced for inattentive respondents, and we interpret this as a type of anchoring (picking the same number for the second list as for the first list). Given that inattentives choose 0 too often under treatment condition (relative to under control) in list A and exactly the same respondents receive control in list B, they choose 0 too often now under control (relative to under treatment). Also, inattentive respondents under control choose the maximal number more often in list A and now they get the treatment condition. The anchoring leads them to choose the maximal number too often now under the treatment. These are the mechanics behind the reversal seen between Figures 3.4 and 3.5.

Our findings have two important implications. First, our results indicate that responses from the second list in a double list experiment may be constrained by an 'anchoring effect.' This observation implies that inattentiveness may undermine the premises of a double list experiment, overshadowing any efficiency gain. Second, as none or the maximal number of items are chosen disproportionately too often or rarely under different conditions for inattentive respondents, we have identified a violation of the assumption of no design effect. Survey inattention may result in violation of a key assumption underlying the experimental design and may thus undermine the value of these designs.

Since attentive and inattentive respondents also differ demographically, we estimated linear regression models to evaluate whether differential responses persist after controlling for basic individual attributes (see Table B13 in the supplementary materials). According to our most comprehensive specification (Model 3), inattentive respondents on average choose 0.28 fewer items than attentive respondents, holding demographics and other individual characteristics fixed. Respondents who failed a trap question are much less likely to support anti-immigration groups according to list A (by 39% and 11% respectively in absolute magnitude), and much more likely to support anti-immigration groups according to list B (by 44% and 33% respectively).[15] All effects are large in magnitude and are statistically significant except for the 11% decrease for organization Y for list A. The reversal seen here across lists represents strong evidence that inattentiveness is driven more by shirking than by genuine disinterest in politics.

[15] 44% and 33% come from the following calculations: 0.44 = 0.83 – 0.39, 0.33 = 0.44 - 0.11.

3.5 Discussion: Dealing with Inattentiveness

As we have shown, most polls and surveys are likely to contain many inattentive respondents, and they provide different responses to direct and indirect survey questions from attentive respondents. What should we do to deal with survey inattentiveness? Four general approaches for addressing inattentiveness include: (1) doing nothing (keeping all respondents in the sample and ignoring attentiveness in data analyses); (2) dropping respondents flagged as inattentive and analyzing the rest of the data without further adjustment; (3) dropping respondents flagged as inattentive and re-weighting the rest of the data; and (4) keeping all respondents in the sample and accounting for attentiveness via model-based statistical adjustment.

If lack of careful attention to the questionnaire is innocuous, in the sense that it does not alter responses to survey questions, then the best approach for dealing with inattentiveness is to do nothing. This is reasonable in situations where inattentiveness is associated with a lack of interest in politics, provided that uninterested respondents answer similarly regardless of the amount of attention to the questionnaire and concentration effort. In such cases, behaviors such as selecting "don't know" or indicating indifference between response alternatives constitute genuine reflections of inattentives' attitudes toward politics, and should therefore not require adjustment.

If inattentives report different answers than they would were they paying attention—as could be the case for respondents that engage in satisficing behavior for reasons other than lack of interest in politics—then doing nothing is not reasonable, as it could lead to inaccurate inferences. Oppenheimer, Meyvis, and Davidenko, 2009 instead argue that "eliminating participants who are answering randomly . . . will increase the signal-to-noise ratio, and in turn increase statistical power." Dropping inattentives, the second option, improves efficiency by reducing noise. However, depending on the study and particularly the subject pool, attention may be correlated with individual characteristics, such as age, gender, and education. This is not the case for Oppenheimer, Meyvis, and Davidenko, 2009 but is the case for Berinsky, Margolis, and Sances, 2014, and our study. For the latter case, if these measured and unmeasured individual characteristics are important correlates of the outcome of interest, then simple elimination of inattentive respondents would lead to a sample that is not representative of the target population (Berinsky, Margolis, and Sances, 2014).

An alternative approach in these situations is to drop inattentive respondents from the analysis and re-weight the sample to obtain population estimates. If the weighting scheme adequately accounts for the probability of inclusion in the sample being inversely related to correlates of inattentiveness—such as being young and having low levels of education—then this approach could help correct for sampling bias. The performance of re-weighting depends on what inattentives' responses would be, were they to pay attention. If these counterfactual responses are close to those given by attentives with similar individual characteristics, then this approach recovers the true population quantities of interest. By training inattentive respondents, Oppenheimer, Meyvis, and Davidenko, 2009 find that forcing respondents who fail IMCs to try again until they pass converts inattentive to attentive respondents. However, Berinsky, Margolis, and Sances, 2014 were unable to replicate this finding suggesting that more research is necessary to determine inattentives' counterfactual responses.

Dropping inattentives and re-weighting the sample, however, also presents a few limitations. First, it assumes that analysts have access to valid and reliable measures of attentiveness that they can use in deciding whether to keep or drop respondents. Since respondents' concentration efforts are not directly observed, attentiveness is likely to be measured with error. This is supported by the work by Berinsky, Margolis, and Sances, 2014 that finds that there is not a perfect correlation across IMCs both within and across surveys. Moreover, polar cases of complete absence or presence of attention are likely to be rare; therefore, deciding the minimum level of attention necessary for keeping respondents in the sample may be far from straightforward. Second, inattentiveness may depend on individual characteristics associated with the outcome of interest. In that case, dropping inattentives from the sample may alter estimates, a problem analogous to using listwise deletion for handling cases of data missing not at random (Pepinsky, 2018). For example, if inattentives genuinely have more moderate opinions on policy issues than attentives, then dropping inattentives could lead analyst to infer that public opinion is more polarized than it actually is (R. Michael Alvarez and Franklin, 1994; R. Michael Alvarez, 1997). Third, dropping inattentives would exacerbate existing unit nonresponse problems and require analysts to rely more heavily on survey weights than they otherwise would. Adjusting survey weights as to account for unit non-response due to inattentiveness would require access to auxiliary variables predictive of inclusion in the sample (i.e. of propensity to provide attentive responses), information that may not always be observable or available to practitioners (Bailey, 2017). And lastly, this approach does not allow evaluating the ways inattentiveness distorts answers to survey questions, which might be of substantive interest to some researchers.

A fourth approach is to develop a statistical model relating outcome variables to measure(s) of attentiveness, controlling for individual attributes that may be associated with both the attention paid to the questionnaire and the attitude or behavior of interest. In the absence of systematic error in the measure of attentiveness, this model-based approach could be used to learn about the relationship between inattentiveness and expected values of the outcome variable. When multiple indicators of attentiveness are available in the data set, analysts may be able to incorporate information about measurement error associated with attention assessments into the analysis, which would lead to more accurate estimates of uncertainty about quantities of interest (e.g. standard errors accounting for uncertainty in the attention assessment).

Ultimately, researchers must weigh the benefits of measuring respondents' attention to the questionnaire and adjusting for inattentiveness, against the costs of doing so. In the case of the survey analyzed in this paper, for instance, the polling firm recommended the inclusion of attention filters and did not charge for incomplete responses from respondents that failed trap questions. Because of budget constraints, a decision was made to follow this advice and filter out inattentive respondents. Collecting measures of attentiveness via attention checks also requires lengthier questionnaires and increased administration times. It may also have other consequences including the inducement of Hawthorne effects by motivating participants to provide socially desirable responses or to censor their responses because of fears that anonymity has been lost (Clifford and Jerit, 2015; Vannette, 2017). But not including attention checks (or failing to collect auxiliary information on attention) prevents the researcher from assessing the influence of inattentiveness on study findings and conclusions. If researchers want to ensure a minimum number of considerate responses and the cost per response is not adjusted for attentiveness, then keeping inattentive respondents in the sample may further increase overall costs.

An strategy that may reduce the financial cost of surveys to researchers is placing attention checks throughout the questionnaire (these could be simple instructed-response items or more complex IMCs, depending on technical and time restrictions, as well as types of questions used to measure variables of interest); using these checks to measure attentiveness in combination with collection of metadata such as response time (which typically can be recorded for free); and then negotiating a lower cost per response on account of the number of seemingly inattentive respondents. Simple criteria could be used in the negotiation with the polling firm, such as only counting—for the purpose of determining whether the designated sample size has been reached—responses from individuals that complete the survey within a reasonable amount of time. What constitutes a reasonable response duration can be determined by the researcher while pilot-testing the online questionnaire or during the soft launch of the survey, by looking at the relationship between total time spent completing the questionnaire and attentiveness measured based on ability to pass trap questions. In implementing this recommendation researchers should make sure to ask the polling firm to record all responses, including those given by respondents completing the survey within less than the designated time minimum. Subsequently, summary information on respondent attentiveness can be incorporated into analyses of attitudes and behavior reported by the entire sample of respondents, using statistical techniques suitable for the data and research question at hand.

3.6 Conclusion

Using data from a recent online survey that included trap questions, we evaluated the prevalence and implications of survey inattentiveness. Our results show that many respondents pay little attention to survey questions in self-completion surveys. Younger and less educated respondents, in particular, are more likely to fail trap questions. Inattentives exhibit many of the symptoms of survey satisficing, including speeding and higher frequency of 'don't know' responses. We also studied whether ignoring respondent attentiveness may lead to a biased evaluation of the incidence of critical attitudes and behavior. We found that when asked directly about attitudes and behavior, inattentives provide lower-intensity responses; this is also the case when they are interrogated indirectly about sensitive issues. The results of our double list experiment suggest that inattentiveness is associated with shifts in the propensity to select sensitive items and that the presence of inattentives could challenge fundamental assumptions underlying the experimental design. On the whole, these results show that ignoring inattentive survey respondents risks significant biases in attitudinal and behavorial models.

We argue that researchers should take attentiveness seriously in survey-based studies of political behavior. In the end, what to do with inattentive survey respondents comes down to a question of survey costs relative to survey errors. Evaluating respondent attentiveness using attention checks comes at a cost. The increased questionnaire length and completion time, caused by the addition of survey items, may lead to greater respondent fatigue, administration expenses, and may influence responses to later question (Anduiza and Galais, 2016). On the other hand, while preventing inattentives from completing the survey may reduce noise and bring down the cost of administering a survey, this may make subsequent analysis of these samples more complicated as they may require reweighting or other types of statistical adjustment to enable population-level inferences.

It may also be possible to learn about attentiveness by looking at survey metadata and response patterns, including the time it takes respondents to answer specific questions or to go over the entire questionnaire, as well as by examining the frequency of straightlining and tendency to report non-attitudes or unreasonable responses. More research needs to be done to assess the extent to which alternative indicators provide complimentary information about attentiveness and develop methods to combine information from numerous indicators—including different types of trap questions, varying in terms of difficulty and type of challenge—into useful indicators of overall attentiveness, and to establish guidelines for incorporating this information into standard data analyses.

Chapter 4

SURVEY ATTENTION AND SELF-REPORTED POLITICAL BEHAVIOR

Alvarez, R. Michael and Yimeng Li (2021). “Survey Attention and Self-Reported Political Behavior”. en. In: *APSA Preprints*. DOI: 10.33774/apsa-2021-x689s. URL: https://preprints.apsanet.org/engage/apsa/article-details/61a901af704d057d023da5cf.

4.1 Introduction

Survey research has been in the midst of vast changes in recent years, as new technologies provide new opportunities. New data sources can improve sampling and survey inference (Green and Gerber, 2006; M. Barber et al., 2014; Ghitza and Gelman, 2020), and researchers can contact and interview respondents using many different modes (Atkeson, Adams, and R. Michael Alvarez, 2014). New and sophisticated methods for weighting survey data are now available, helping researchers deal with the complexity of survey sampling and design (Caughey, Berinskey, et al., 2020). And innovative new analytical methods allow researchers to use millions of survey responses, measured across many decades, to analyze opinion change (Caughey and Warshaw, 2018).

One of the primary issues is how these methodological changes might affect the quality of the survey response, which has been the subject of numerous recent studies (Meade and S. B. Craig, 2012; Ansolabehere and Brian F. Schaffner, 2014; Atkeson, Adams, and R. Michael Alvarez, 2014; Maniaci and Rogge, 2014). Of particular interest has been whether the presence of an interviewer (say in a live telephone survey, or with an in-person interview) changes the incentives for survey respondents to pay close attention to survey questions and to provide accurate answers, which has been studied for decades by survey methodologists (Bradburn and Sudman, 1974). For example, survey methodologists have studied how interviewers may affect responses to certain types of survey questions, and whether respondents are more likely to provide more accurate information regarding controversial or sensitive questions when there is no interviewer present, especially for self-completion surveys (Tourangeau and Smith, 1996).

But with the increasing use of self-completion surveys, the absence of an interviewer may mean that respondents could move through a questionnaire quickly, and not pay close attention to the questions or the potential responses (Curran, 2016). Inattentive respondents in self-completion surveys may thus provide lower-quality data, as they might randomly answer questions, provide inaccurate responses, answer with a "don't know", or use other tactics to get through a survey quickly (Atkeson, Adams, and R. Michael Alvarez, 2014; Clifford and Jerit, 2015). These concerns have led researchers to study the use of "attention checks", "trap questions", "screeners", usually in the form of instructed response items or instructional manipulation checks (in this paper, we use the term attention checks).

There have been a number of recent studies that have examined the use of attention checks in surveys and opinion polls, documenting how attentive and inattentive survey respondents differ, studying different types of attention checks and methods for measuring attentiveness, and examining how to deal with inattentive respondents in survey data (Read, Wolters, and Berinsky, 2021; Berinsky, Margolis, Sances, and Warshaw, 2021; Kung, Kwok, and Brown, 2018; D. J. Hauser and Schwarz, 2015; Berinsky, Margolis, and Sances, 2014; Oppenheimer, Meyvis, and Davidenko, 2009). There have also been studies that have looked at experimental subject attentiveness, in either convenience samples of students or from crowd-sourcing applications like MTurk (Ahler, Roush, and Sood, 2021; Curran and K. A. Hauser, 2019; Curran, 2016; Thomas and Clifford, 2017). However, the lack of ground truth information prevents researchers from quantifying the performance of attention checks and comparing different approaches to dealing with inattentive respondents. In our paper, we attempt to fill this gap by examining responses to factual survey questions that we can validate with external administrative data.

Our results indicate that respondents failing attention checks are more likely to misreport various factual information. Many of these inattentive respondents nonetheless provide responses in line with the information in the administrative records. For turnout histories in recent elections, which correlate with respondent attention, dropping inattentive respondents leads to an unrepresentative subsample and, hence, estimates with larger biases and variances. By contrast, for modes of voting in recent elections, which are largely uncorrelated with attention check passages, dropping inattentive respondents yields estimates with smaller biases that often outweigh the cost of larger variances.

In the next section we discuss theory and past research, then the unique dataset and methodology that we use to test hypotheses about inattentive survey respondents.

We next present results from our analysis, and conclude with a discussion of what our results imply for survey research.

4.2 Theory and Past Research

Inattentive respondents may diminish the quality of survey data. For example, one recent study found that inattentive respondents offer quick answers, are more like to give “don’t know” answers, and are more likely to “straightline” (R Michael Alvarez et al., 2019). Theoretically, survey respondents often may engage in satisficing — which can occur when individuals encounter cognitively challenging tasks, and they do not expend the resources necessary to fully comprehend the question, to search their memory for the best answer, or otherwise provide only superficial attention to a survey question (Simon, 1956; Jon A Krosnick, 1991).

One solution to the problems caused by inattentive respondents is to use attention checks, also known as “screeners” or “trap questions”. Past research has differentiated attention checks into two types. One type of attention check is an instructional manipulation check (IMC), where there is a deliberate change in the instructions in a survey question designed to capture whether the respondent is reading and cognitively processing the question’s instructions (Oppenheimer, Meyvis, and Davidenko, 2009; Berinsky, Margolis, and Sances, 2014; Anduiza and Galais, 2017). An example of an IMC is adding a clause to a survey question instructing the respondent to ignore the question and provide a specific answer. The other type of attention check is an instructed response item (IRI), where the responses to a survey question are altered in a way that should elicit whether the respondent is attentive to the question’s response options (R Michael Alvarez et al., 2019). An example of an IRI is adding a row in a grid instructing respondents to select ‘strongly disagree’ for survey quality control.

However, due to the lack of ground truth information, previous studies have relied on various post-hoc measures to evaluate the performance of attention checks.[1] These measures look at whether respondents passing and failing attention checks differ in producing canonical experimental results, empirical correlations between negatively correlated survey questions, straight-lining behavior, and response time. For the same reason, it has also been impossible to evaluate the performance of different statistical approaches to dealing with inattentive respondents. Our research

[1] In a recent related study, Kuhn and Vivyan, 2021 compared the accuracy of self-reported behavior using list experiments and direct questions, and comparing the responses to administrative data.

takes advantage of a unique survey, which has questions that we can validate against administrative data, as well as measures of respondent attentiveness.

We motivate our research reported in this paper using the theory of satisficing (Simon, 1956; Jon A Krosnick, 1991). We hypothesize that inattentive survey respondents will provide less accurate answers to questions about their recent and past voting behavior, in particular whether they voted in the current and past elections, and the method that they used to vote in the current and past elections (Hypothesis 1).

Because inattentive respondents provide less accurate answers, that generates error and noise in a survey dataset, so that when the data is used for modeling or testing hypotheses, including the answers from inattentive respondents will potentially generate bias in model estimates and incorrect inferences. Moreover, inattentive survey respondents who are engaging in satisficing will follow particular heuristics when they provide survey responses, like straightlining, midpoint responses, or socially desirable answers. Thus it is likely that inattentive respondents will generate systematic errors in a survey dataset, not just additional noise or variance. This leads to Hypothesis 2: inattentive respondents will provide answers that can introduce bias in model estimates.

The implications of our analysis are important for dealing with inattentive respondents in research using self-administered surveys. We find consistent support for Hypothesis 1, for self-reports of turnout for a number of recent elections, and for self-reports of the method that the voter used to return or cast their ballot. We also show support for Hypothesis 2, that there is evidence that inattentive respondents are generating systematic error in survey data (not random error). Thus respondent inattention will generate error and noise in survey self-reports of political behavior, and under the assumption that respondent attention is uncorrelated with the outcome of interest, we argue that dropping inattentives from the analysis will increase variance because of information loss, but will decrease bias in the estimates. However, under the assumption that respondent attention is correlated with the outcome of interest, if a researcher drops inattentive respondents that can produce bias by creating an unrepresentative sample (though the direction of the bias is not clear, and will depend on the type of heuristic that inattentive respondents use in answering the survey). Removing inattentive respondents in this situation will also increase variance.

Testing these hypotheses with ground-truth data is the gap in the literature that our research seeks to fill. In 2018 we fielded a unique survey that allowed us to connect survey responses to voter registration data. In our survey, we posed a series of questions, regarding current and past electoral participation as well as other demographic information, that we could validate with the administrative data. This provides us with a larger array of different types of ground-truth information, so that we can study the accuracy of inattentive and attentive survey respondents.

In the next section of our paper, discuss our survey design. Then we turn to various tests of our two hypotheses using data from our unique survey. We first examine how inattentive respondents answer questions about their voting participation and method of voting. We then examine our second hypothesis, and test for whether inattentive respondents are generating answers that introduce bias into models using their survey responses.

4.3 Survey and Attention Checks

Our study uses voter registration and voting history administrative data provided by the Orange County Registrar of Voters (OCROV). These datasets contain information about each registered voter in the county, including contact and demographic information, partisan registration, whether they turned out to vote in past elections, and if they did turn out in a past election whether they voted in person or by mail. Importantly, in California, voters can provide contact information (like a telephone number or email address) when they register to vote. Of the approximately 1.6 million voter registration records in the county in the 2018 general election, over 530,000 of those records were associated with an email address. We used all of the records with email addresses for our survey.

For our research on the 2018 general election in Orange County, we developed a self-completion online survey focused on the voter's election experiences. The online questionnaire included attention checks in the form of IMCs and IRIs, questions on voter registration, turnout in recent elections, and method of voting, in addition to questions on other subjects. We invited registered voters (via email) to participate in our survey between Thursday, November 8, 2018, and Tuesday, November 13, 2018. From 531,777 invites to all registered voters with email addresses, we received 6,952 complete responses. The survey took about 12 to 15 minutes to complete (the median duration was 13 minutes) and was provided in English. More details about the survey can be found in the paper's Supplementary Materials.

We then match back each survey response to the corresponding registered voter. In most cases (6,816), the survey respondent can be linked back to a unique registered voter by the email address alone.[2] In cases of ambiguity, we further match responses to voters according to age (or, in rare cases, a combination of age and gender) from self-reports and administrative records. By matching survey respondents directly to administrative data, we can validate self-reports of voter turnout and method of voting and add features (such as party registration) from the administrative data to our analysis. More details about the administrative data are discussed in the paper's Supplementary Materials.

Our 2018 survey contained both Instructional Manipulation Checks (IMC) and Instructed Response Items (IRI) for assessing survey attention.[3] We designed the survey instrument so that respondents would receive one attention check approximately 25% of the way through the survey (attention check 1), the second when they had completed about 50% of their survey (attention check 2), and the final attention check at about 75% completed (attention check 3). In all cases, subjects who ignored the attention checks were allowed to continue to the next set of survey questions.

To avoid potential question order biases, we randomized the appearance of the attention check questions in our survey. For attention check 1, respondents were randomly assigned to receive either an IRI or IMC attention check. The IRI check asked subjects to answer "oppose" or "support" among five other questions about their opinions regarding election reform; the IMC attention check asked subjects to ignore a question on news media websites and to select two specific news organizations as their answers.

Attention check 2 followed a series of questions about the subject's voting experiences. Those who had been asked to answer the IRI for attention check 1 were then asked the IMC for attention check 2. Similarly, subjects who answered the IMC for attention check 1 were asked the IRI attention check. Again, we designed this to avoid any question order or location effects with respect to the use of the IMC or IRI format as attention checks.

Finally, the third attention check was located about 75% of the way through the survey questionnaire. Here the subjects were again randomized, with half receiving

[2]For each email address in the voter registration file, we generated a unique survey URL. This allows us to link the survey responses to records in the voter registration file when there is a unique record associated with the email address (6,816 out of 6,952 cases).

[3]We used an IMC studied in Berinsky, Margolis, and Sances, 2014 and an IRI studied in R Michael Alvarez et al., 2019 as our main attention checks.

Table 4.1: IMCs Screen Respondents More Aggressively Than IRIs

	Fail	Pass	Skip
IRI - Attention check 1 or 2	15%	82%	4%
IRI - Attention check 1	16%	81%	4%
IRI - Attention check 2	13%	83%	4%
Additional IRI	9%	88%	4%
IMC - Attention check 1 or 2	45%	52%	3%
IMC - Attention check 1	49%	47%	4%
IMC - Attention check 2	40%	57%	3%
Additional IMC	28%	70%	2%

an additional IRI, while the others received an additional IMC. We included this attention check (an additional IRI or an additional IMC) for the purposes of robustness checks; we present our results from the main IRI and IMC attention checks in the text.[4]

Questions on turnout and mode of voting in the 2018 General Election appear right before attention check 1, while questions on turnout and mode of voting in previous elections appear right after attention check 2. We provide the complete wording of the attention checks and these questions in the paper's Supplementary Materials.

The overall passage rates of our IRI and IMC attention checks are presented in Table 4.1. The first four rows of the Table provide statistics on failure rates for the IRI questions, while the last four rows show similar statistics on the failure rates for the IMC questions. Consistent with previous studies on self-administered online surveys (Berinsky, Margolis, and Sances, 2014; R Michael Alvarez et al., 2019), a significant fraction of respondents failed these attention checks. As we expected, IMCs screen respondents more aggressively than IRIs, with 45% of respondents failing the IMC at either attention check 1 or 2 and 15% of respondents failing the IRI at either attention check 1 or 2.

Also, it is important to note two other results in Table 4.1. First, regardless of whether a respondent receives the IRI or the IMC question at attention check 2 (after they have had one of the attention checks earlier in the survey), we see that the passage rates for both are slightly higher later in the survey. Similarly, for the third attention check, again, passage rates are higher as well for both types of attention

[4] In the third attention check, the IRI asked subjects to answer either "disagree strongly" or "agree strongly" after a battery of questions about political polarization. The final attention check's IMC involved a question about social media; subjects were asked to ignore the question and select two specific answers to show they are paying attention to the survey questions.

check questions. This indicates that as subjects proceed through the survey, the presence of successive attention checks may increase the cognitive focus of survey respondents.[5]

Furthermore, the passage of our IRI and IMC attention checks is correlated with important demographics, as shown in Table C.1 in the Supplementary Materials. Educational attainment is positively correlated with the passage of each type of attention check, consistent with previous studies on self-administered online surveys (Berinsky, Margolis, and Sances, 2014; R Michael Alvarez et al., 2019). Non-white respondents are less likely to pass IRI and IMC attention checks than white respondents, again consistent with previous research. Male and female voters appear equally likely to pass the attention checks. Finally, contrary to previous studies, we find that age is negatively correlated with the passage of each type of attention check.

4.4 Results

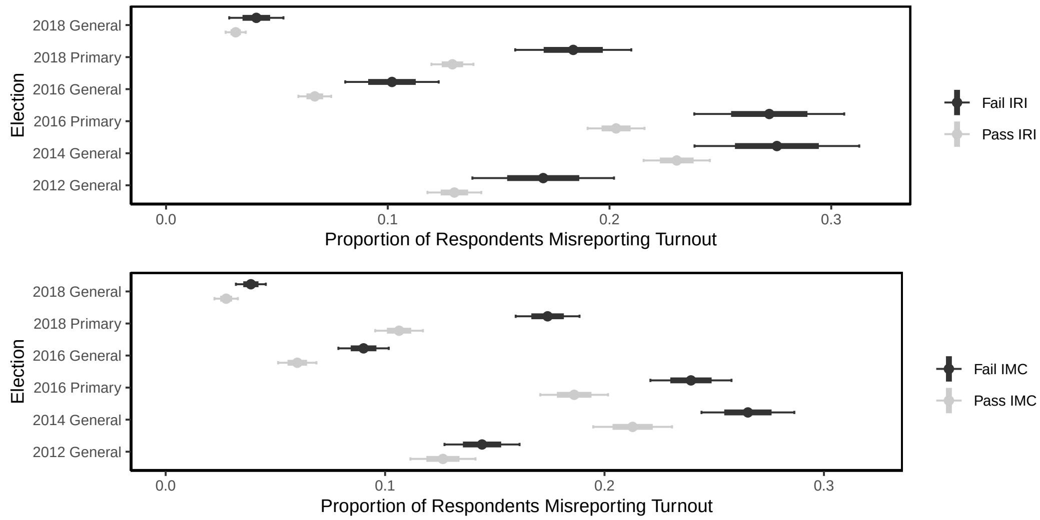
Before turning to our main results, we first look at three pieces of basic respondent information that we are able to validate using the administrative records (presented in Table C.2 in the Supplementary Materials): year of birth, city of residence, and voter registration before or after January 1, 2017, For each of the three pieces of basic respondent information and each of the two types of attention checks, respondents failing either the IRI or IMC attention check are more likely to misreport the information than those passing either attention check.

We begin by testing Hypothesis 1, that inattentive survey respondents will provide less accurate answers to questions about their recent and past voting behavior. Here we focus on whether they voted in the current and past elections, and the method that they used to vote in the current and past elections. Our first step is to compare the accuracy of self-reported turnout between respondents passing and failing each type of attention check, as shown in Figure 4.1. The analysis reported in Figure 4.1 takes advantage of the fact that we can ask respondents for self-reports about whether they participated in a number of previous elections, and that we can then check the veracity of their self-reports against our administrative data.

It's important to note that we consistently see over-reporting of turnout in our data. But for the purposes of testing Hypothesis 1, the immediate observation from Figure 4.1 is that respondents failing either the IRI or the IMC attention check are

[5] These results require further and future study, but are beyond the scope of the research we report in this paper.

Figure 4.1: Inattentive Respondents Are More Likely to Misreport Turnout

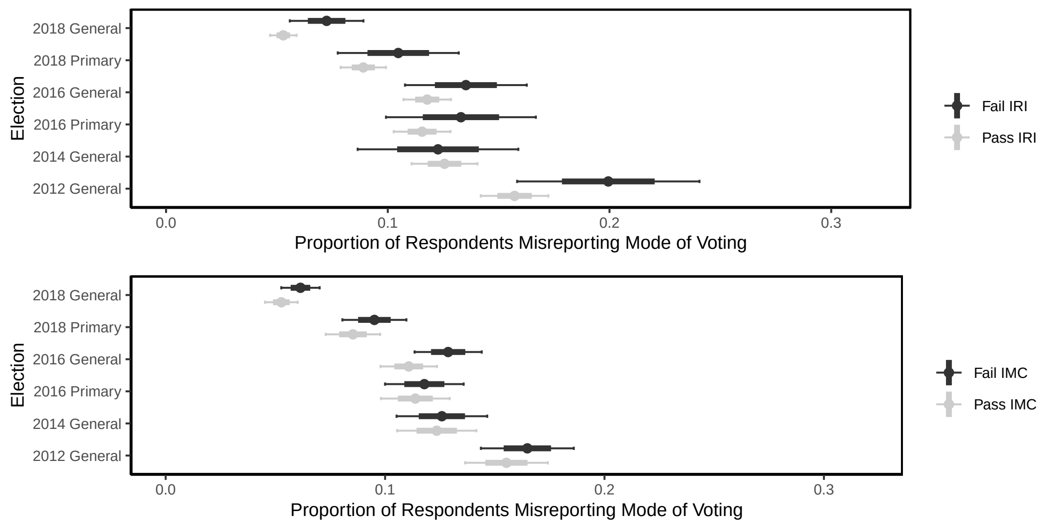


Note: This figure displays the proportion of respondents misreporting turnout in six recent elections among those who failed the IRI (top panel, black), passed the IRI (top panel, grey), failed the IMC (bottom panel, black), or passed the IMC (bottom panel, grey). In each case, dots display the point estimates, thin intervals display 95% confidence intervals, and thick intervals display 90% confidence intervals.

significantly more likely to misreport turnout than those passing either attention check. While the magnitude of the difference is modest for the reporting of turnout in the 2018 general election (a few days before the survey), the difference is generally sizable for many past elections. This is not surprising, because trying to remember participation in past elections requires greater cognitive effort (especially primary elections, which are less salient compared to general elections), as subjects need to search their memory more extensively in order to provide a correct report. For example, 18.4% and 17.4% of respondents failing the IRI and IMC, respectively, misreported turnout in the 2018 Primary Election, compared to 12.9% and 10.6% who passed these attention checks.

These results indicate that including inattentive respondents as measured by these attention checks will lead to bias in the estimation of turnout. Also evident from Figure 4.1, however, is that many respondents who failed the attention checks but provided accurate self-reports of turnout nonetheless. This observation implies that dropping inattentive respondents as measured by these attention checks will increase the variance in the estimation of turnout. Finally, notice that respondents passing IMC provided the most accurate account of turnout, followed by respondents passing IRI, and then all respondents. This observation is consistent with the pattern that

Figure 4.2: Inattentive Respondents Are More Likely to Misreport Mode of Voting



Note: This figure displays the proportion of respondents misreporting mode of voting in six recent elections among those who failed the IRI (top panel, black), passed the IRI (top panel, grey), failed the IMC (bottom panel, black), or passed the IMC (bottom panel, grey). In each case, dots display the point estimates, thin intervals display 95% confidence intervals, and thick intervals display 90% confidence intervals.

IMC screens out respondents more aggressively than IRI.

Our administrative data also contains information about the method that each voter used to cast their ballot in past elections, and in our survey, we asked each respondent to recall their method of voting. Our second test of Hypothesis 1 examines the difference in terms of accurately reporting the method of voting between respondents passing and failing each type of attention check, with results presented in Figure 4.2. Again, here we see consistent misreporting in our self-reported survey data. While the differences seen in this figure are less pronounced than the difference shown in Figure 4.1 for the accuracy of reported turnout, respondents failing IRIs are significantly more likely to misreport mode of voting than those passing IRIs, in all of the elections except the 2014 General Election. The same is true for IMC, but only significant for more recent elections.

One interesting pattern seen in Figure 4.2 is that generally the extent of misreporting is greater for elections further in the past, than for the most recent election. This pattern suggests that recall of the method that a voter used to obtain and return their ballot might be cognitively demanding for voters, in particular some of them who are not paying close attention as they complete the questionnaire.

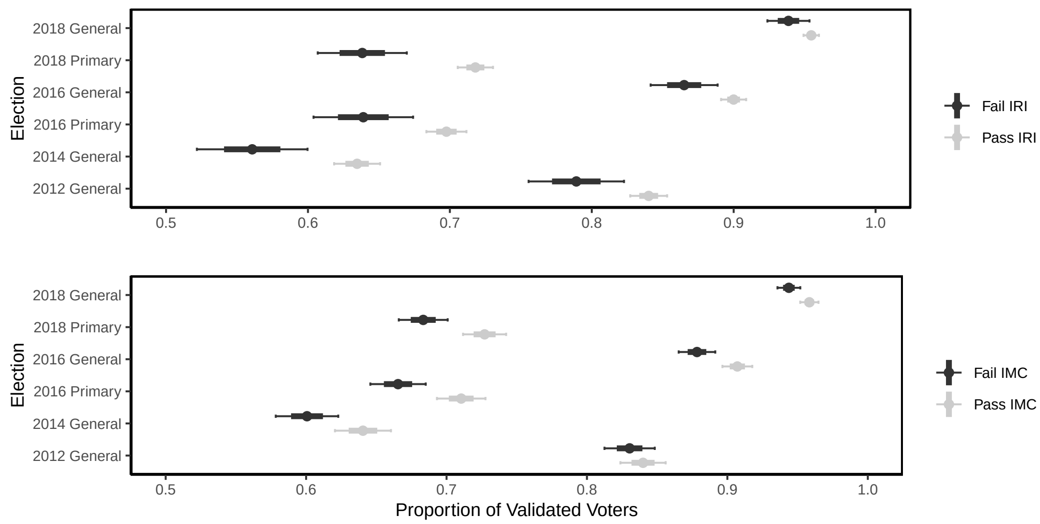
The magnitude of the accuracy difference between self-reports of mode of voting by respondents failing and passing the attention checks averages around 1.8% and 1.0% for IRI and IMC, respectively, for the past two election cycles. This result indicates that including inattentive respondents as measured by these attention checks will lead to bias in the estimation of mode of voting, modest but statistically significant in most cases. As is the case with turnout, we find that many voters reported mode of voting accurately regardless of attentiveness. This pattern suggests that dropping inattentive respondents will lead to less precise estimates of mode of voting. Ultimately, the bias-variance trade-off will govern the best strategy to deal with inattentive respondents, which we shall explore in greater detail later.

So far, we have used validated survey responses to establish that inattentive respondents provided less accurate reports of turnout and method of voting, both for contemporary and past elections. Our findings are consistent with previous research that found inattentive and attentive respondents provide different self-reports of key political behavior and attitudes (for example, R Michael Alvarez et al. 2019). While these results highlight the bias in the estimation of political behavior if respondent inattention is not accounted for, dealing with inattentive respondents is complicated by the fact that respondent attention may be correlated with political behavior and attitudes of interest.

Our second hypothesis regards this issue, that inattentive respondents will provide answers that can introduce bias in model estimates. We test our second hypothesis and its corollaries in a number of ways in Figures 4.3- 4.6. First, Figure 4.3 presents the correlation between respondent attention and turnout by comparing validated turnout between respondents passing and failing each type of attention check. Clearly, respondents failing the attention check are significantly less likely to turn out to vote than those passing the attention check, for both IRI and IMC. The magnitude of the difference ranges from 1.6% (2018 General) to 8.0% (2018 Primary) for IRI and 1.0% (2012 General) to 4.5% (2016 Primary). Our result is the first to document the positive correlation between respondent attention and political participation using validated turnout.[6] The implication is that dropping inattentive respondents, a common way to address respondent inattention, can introduce a source of bias by creating a sample unrepresentative of the population in terms of political participation. Again, consistent with the fact that the IMC screens out

[6] Alvarez et al. (2019) found positive correlations between respondent attention and self-reported political participation.

Figure 4.3: Respondent Attention Is Positively Correlated with Validated Turnout



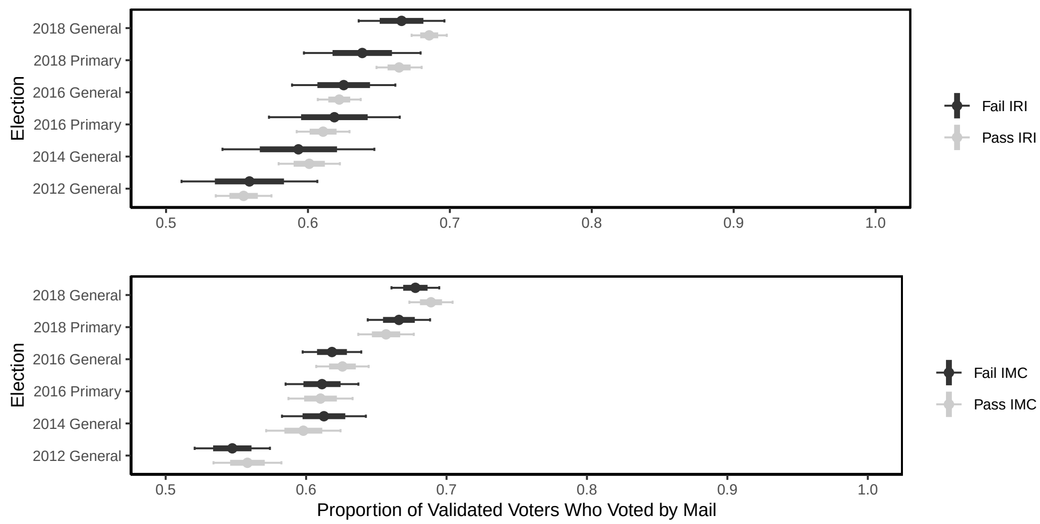
Note: This figure displays the proportion of respondents confirmed to have participated in six recent elections among those who failed the IRI (top panel, black), passed the IRI (top panel, grey), failed the IMC (bottom panel, black), or passed the IMC (bottom panel, grey). In each case, dots display the point estimates, thin intervals display 95% confidence intervals, and thick intervals display 90% confidence intervals.

respondents more aggressively than the IRI, respondents passing the IMC had the highest level of turnout, followed by the IRI, and then all respondents.

We next examine the correlation between respondent attention and mode of voting by comparing validated mode of voting between respondents passing and failing each type of attention check in Figure 4.4. In contrast with turnout, respondents failing the attention check are not different from those passing the check in terms of their actual choice of voting method. This result indicates that dropping inattentive respondents will not introduce bias that comes from an unrepresentative sample in terms of mode of voting. As we discuss in detail below, the presence or absence of correlation between respondent attention and political behavior or attitudes of interest dictates the magnitude of this important source of bias and ultimately factors heavily into the consideration of strategies to deal with respondent inattention.

With these results in hand, we can now work toward developing the best strategy to deal with inattentive respondents in different scenarios. To illustrate the considerations going into such decisions, we consider the simplest possible strategy, dropping inattentive respondents identified by the attention checks. We first compare the performance of turnout estimates based on all respondents, respondents passing the IRI, and respondents passing the IMC, in terms of bias, standard error,

Figure 4.4: Respondent Attention Is *Not* Correlated with Validated Mode of Voting

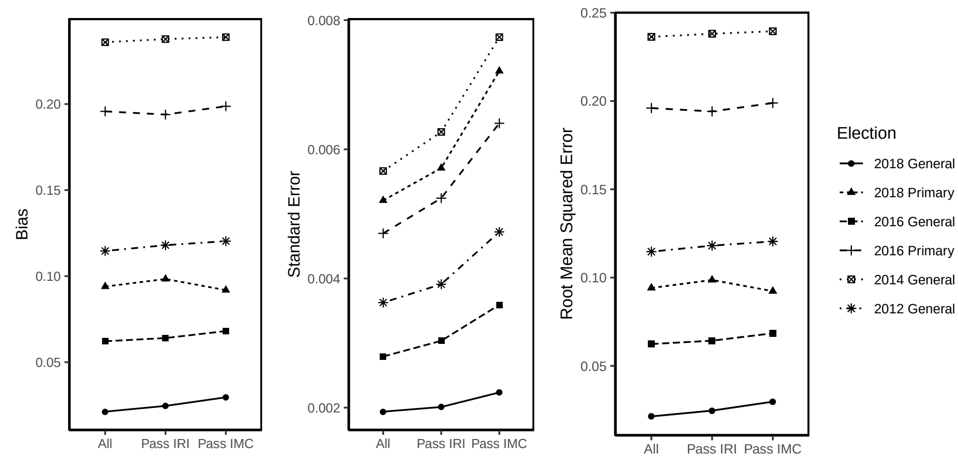


Note: This figure displays the proportion of respondents confirmed to have voted by mail in six recent elections among those who failed the IRI (top panel, black), passed the IRI (top panel, grey), failed the IMC (bottom panel, black), or passed the IMC (bottom panel, grey). In each case, dots display the point estimates, thin intervals display 95% confidence intervals, and thick intervals display 90% confidence intervals.

and root mean squared error, as shown in Figure 4.5. While respondents failing the attention checks are significantly more likely to misreport turnout (Figure 4.1), the reduction in bias from dropping these inattentive respondents is countervailed by the increase in bias from the unrepresentativeness of attentive respondents in terms of turnout (Figure 4.3). As a result, dropping inattentive respondents does not reduce bias in turnout estimates (left panel). Respondents failing the attention checks, moreover, often provide correct self-reports of turnout nonetheless, leading to the pattern that keeping all respondents yields turnout estimates of the smallest standard errors (middle panel). Given the patterns present in terms of bias and variance, it's unsurprising that we find that the turnout estimates based on all respondents have the smallest root mean squared errors in most cases (right panel).

We next look at the difference in the performance of estimates of proportions of by-mail voters based on all respondents as well as respondents passing each type of attention check, with results presented in Figure 4.6. In contrast to the results for turnout, since respondent attention as measured by attention checks is uncorrelated with their mode of voting, dropping inattentive respondents does not create an unrepresentative sub-sample that would lead to an increase in bias. Dropping respondents failing attention checks, therefore, leads to a reduction of

Figure 4.5: Dropping Inattentive Respondents Does *Not* Reduce Bias in Turnout Estimates



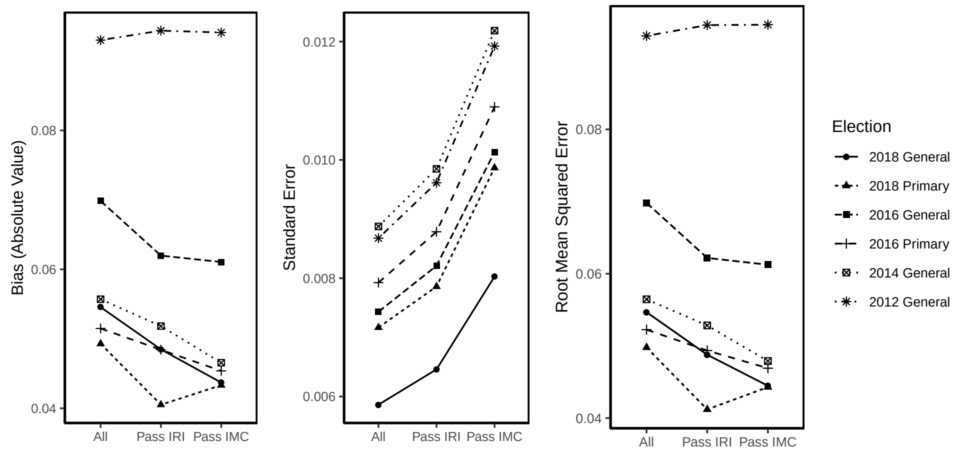
Note: This figure displays the bias (left panel), standard error (middle panel), and root mean squared error (right panel) of turnout estimates based on all respondents, respondents passing the IRI, and respondents passing the IMC, for six recent elections.

bias (left panel, with the sole exception of 2012 General) as these respondents are more likely to misreport their mode of voting. While such procedures still increase imprecision of the estimates compared to keeping all respondents (middle panel), its effect is dwarfed by the reduction of bias. Dropping inattentive respondents, as a result, yield smaller root mean squared errors in estimates of proportions of by-mail voters (right panel). Notice that dropping inattentive respondents reduce bias due to the higher likelihood of misreporting from these respondents and increase variance due to a reduced sample size in both turnout estimates and the estimates of proportions of by-mail voters. The difference in the best strategies for these two quantities of interest comes from the correlation, or the lack thereof, between respondent attention and outcomes of interest, or equivalently, the representativeness of inattentive respondents for the entire sample in terms of the variable of interest.

4.5 Discussion and Conclusion

The implications of our analysis are important for dealing with inattentive respondents in research using self-administered surveys. We find consistent support for Hypothesis 1, that attentive survey subjects are more likely to provide accurate survey responses, for self-reports of turnout for a number of recent elections, and for self-reports of the method that the voter used to return or cast their ballot. We also show support for Hypothesis 2, that there is evidence that inattentive respondents

Figure 4.6: Dropping Inattentive Respondents Reduces Bias in Voting-by-Mail Estimates



Note: This figure displays the bias (left panel), standard error (middle panel), and root mean squared error (right panel) of estimates of proportions of by-mail voters based on all respondents, respondents passing the IRI, and respondents passing the IMC, for six recent elections.

are generating systematic (not random) error in survey data.

We obtained these results using a large sample from a diverse population, and our findings are consistent with the results in studies that have focused on attentiveness in small samples (for example, student or MTurk studies). Our paper, meanwhile, advances survey research in this area by validating survey responses with external administrative data at the individual level, and is the first paper to do so to the best of our knowledge. Thus we argue that our results have general and important implications for survey research, in particular for large-scale surveys and polls. Our focus on specific factual questions (voter registration, turnout, mode of voting) is driven by our ability to validate the responses to these questions in our administrative data. We leave to future research to explore other factual questions that can be validated using other forms of external ground-truth information.

However, the question that we are left with, since we know that there are inattentive subjects in survey studies, is what should a researcher do? One analytic strategy is to simply drop the inattentive subjects from any analysis of survey data. This approach often is implemented during the course of survey research, when subjects who do not pass a particular attention check are not allowed to continue with the remainder of the survey. As we have argued, based on our research, this approach is problematic. If under the assumption that respondent attention is uncorrelated with the outcome of interest, dropping inattentives will increase variance because of information loss, but will decrease bias in the estimates. But if respondent attention is correlated with the outcome of interest, dropping inattentive respondents can produce an unrepresentative sample, and thus bias estimates. While the direction of the bias is not clear, and will depend on the type of heuristic that inattentive respondents use in answering the survey, bias will be an issue in these situations. As it is often not possible to know which assumption is valid, applied researchers should proceed with caution regarding how they deal with inattentive respondents.

Our advice is that survey researchers should avoid dropping inattentive respondents from their analyses, as either increasing variance or producing bias should be avoided. Rather, researchers should use methods that allow them to incorporate the heterogeneity in survey response generated by inattentive respondents—which, if done correctly, can avoid increasing variance and producing bias. Minimally, researchers should test results of survey research for robustness with respect to survey attentiveness, and if there is evidence of potential bias, then include measures of survey attention (either from direct questions or proxies) in models of survey responses. More research is necessary to help develop best practices for measuring attention and for modeling the heterogeneity induced by differences in attention.

Our research design also allows us to compare different types of attention checks. Between instructed response items and instructional manipulation checks, two commonly used types of attention checks, we find that instructional manipulation checks screen respondents more aggressively than instructed response items. As a result, while using instructional manipulation checks as screeners further reduces biases compared to instructed response items in many (but not all) cases, it comes at the cost of larger variances for our estimates. For applied researchers, we suggest considering multiple factors. First, while the variance consideration is secondary for a large survey like ours (over 6,900 respondents), it may be a primary consideration for surveys of small or moderate sizes. Secondly, many survey instruments already contain grid/list questions, making instructed response items much less costly than instructional manipulation checks, which require additional standalone questions just for quality control purposes. Lastly, since multiple attention checks are often recommended for online surveys longer than a few minutes, different types of attention checks can be employed to ensure the robustness of results to various levels of respondent attention screening.

BIBLIOGRAPHY

Ahler, Douglas J., Carolyn E. Roush, and Gaurav Sood (2021). “The micro-task market for lemons: data quality on Amazon’s Mechanical Turk”. In: *Political Science Research and Methods*, pp. 1–20. DOI: 10.1017/psrm.2021.57.

Ahlquist, John S. (2017). “List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators”. In: *Political Analysis*, pp. 1–20. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2017.31. URL: https://www.cambridge.org/core/journals/political-analysis/article/list-experiment-design-nonstrategic-respondent-error-and-item-count-technique-estimators/B5726FB76E31168611ED7EB48F8CAA42 (visited on 01/03/2018).

Ahlquist, John S., Kenneth R. Mayer, and Simon Jackman (2014). “Alien Abduction and Voter Impersonation in the 2012 U.S. General Election: Evidence from a Survey List Experiment”. In: *Election Law Journal: Rules, Politics, and Policy* 13.4, pp. 460–475. ISSN: 1533-1296. DOI: 10.1089/elj.2013.0231. URL: http://online.liebertpub.com/doi/abs/10.1089/elj.2013.0231 (visited on 11/28/2017).

Alvarez, R Michael et al. (2019). “Paying attention to inattentive survey respondents”. In: *Political Analysis* 27.2, pp. 145–162.

Alvarez, R. Michael (1997). *Information and Elections*. en. Google-Books-ID: mJNFDwAAQBAJ. University of Michigan Press. ISBN: 978-0-472-02237-3.

Alvarez, R. Michael and John Brehm (2002). *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. en. Google-Books-ID: dWh10zFdjNUC. Princeton University Press. ISBN: 978-0-691-09635-3.

Alvarez, R. Michael and Charles H. Franklin (1994). “Uncertainty and Political Perceptions”. In: *The Journal of Politics* 56.3, pp. 671–688. ISSN: 0022-3816. DOI: 10.2307/2132187. URL: https://www.jstor.org/stable/2132187 (visited on 09/16/2018).

Anduiza, Eva and Carol Galais (2016). “Answering Without Reading: IMCs and Strong Satisficing in Online Surveys”. In: *International Journal of Public Opinion Research* 29.3, pp. 497–519. ISSN: 0954-2892. DOI: 10.1093/ijpor/edw007. URL: https://academic.oup.com/ijpor/article/29/3/497/2669464 (visited on 12/17/2017).

– (2017). “Answering without reading: IMCs and strong satisficing in online surveys”. In: *International Journal of Public Opinion Research* 29.3, pp. 497–519.

Ansolabehere, Stephen and Brian F Schaffner (2018). “Taking the study of political behavior online”. In: *The Oxford handbook of polling and survey methods*. Oxford University Press New York, p. 76.

Ansolabehere, Stephen and Brian F. Schaffner (2014). “Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison”. In: *Political Analysis* 22.3, pp. 285–303.

Aronow, Peter M. et al. (2015). “Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence”. In: *Journal of Survey Statistics and Methodology* 3.1, pp. 43–66. ISSN: 2325-0984. DOI: 10.1093/jssam/smu023. URL: https://academic.oup.com/jssam/article/3/1/43/915561 (visited on 11/28/2017).

Atkeson, Lonna Rae and Alex N. Adams (2018). “Mixing Survey Modes and Its Implications”. In: *Oxford University Handbook on Polling and Polling Methods*. Ed. by Lonna Rae Atkeson and R. Michael Alvarez. Oxford University Press.

Atkeson, Lonna Rae, Alex N. Adams, and R. Michael Alvarez (2014). “Nonresponse and Mode Effects in Self- and Interviewer-Administered Surveys”. In: *Political Analysis* 22.3, pp. 304–320. DOI: 10.1093/pan/mpt049.

Bailey, Michael A. (2017). *Selection Sensitive Survey Design: Moving Beyond Weighting*. San Francisco, CA.

Barber, Larissa K., Christopher M. Barnes, and Kevin D. Carlson (2013). “Random and Systematic Error Effects of Insomnia on Survey Behavior”. en. In: *Organizational Research Methods* 16.4, pp. 616–649. ISSN: 1094-4281, 1552-7425. DOI: 10.1177/1094428113493120. URL: http://journals.sagepub.com/doi/10.1177/1094428113493120 (visited on 06/04/2018).

Barber, M. et al. (2014). “Online Polls and Registration-Based Sampling: A New Method for Pre-Election Polling”. In: *Political Analysis* 22 (3), pp. 321–335. DOI: doi:10.1093/pan/mpt023.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk”. en. In: *Political Analysis* 20.3, pp. 351–368. URL: https://www.cambridge.org/core/journals/political-analysis/article/evaluating-online-labor-markets-for-experimental-research-amazoncoms-mechanical-turk/348F95C0FBCF21C3B37D66EB432F3BA5.

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”. In: *American Journal of Political Science* 58.3, pp. 739–753.

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw (2021). “Using screeners to measure respondent attention on self-administered surveys: Which items and how many?” In: *Political Science Research and Methods* 9.2, pp. 430–437. DOI: 10.1017/psrm.2019.53.

Blair, Graeme and Kosuke Imai (2012). "Statistical Analysis of List Experiments". In: *Political Analysis* 20.1, pp. 47–77. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpr048. URL: https://www.cambridge.org/core/journals/political-analysis/article/div-classtitlestatistical-analysis-of-list-experimentsdiv/6AEE6C9D3AB6DA410D602CB035D5959A (visited on 11/27/2017).

Blair, Graeme, Kosuke Imai, and Jason Lyall (2014). "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan". en. In: *American Journal of Political Science* 58.4, pp. 1043–1063. ISSN: 1540-5907. DOI: 10.1111/ajps.12086. URL: http://onlinelibrary.wiley.com/doi/10.1111/ajps.12086/abstract (visited on 11/27/2017).

Bowling, Nathan A. et al. (2016). "Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality." en. In: *Journal of Personality and Social Psychology* 111.2, pp. 218–229. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/pspp0000085. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/pspp0000085 (visited on 06/03/2018).

Bradburn, Norman M and Seymour Sudman (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine Publishing Company.

Caughey, Devin, Adam J. Berinskey, et al. (2020). *Target Estimation and Adjustment Weighting for Survey Nonresponse and Sampling Bias*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Caughey, Devin and Christopher Warshaw (2018). "Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014". In: *American Political Science Review* 112.2, pp. 249–266. DOI: 10.1017/S0003055417000533.

Clifford, Scott and Jennifer Jerit (2015). "Do attempts to improve respondent attention increase social desirability bias?" In: *Public Opinion Quarterly* 79.3, pp. 790–802.

Coffman, Katherine B., Lucas C. Coffman, and Keith M. Marzilli Ericson (2017). "The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated". In: *Management Science* 63.10, pp. 3168–3186. ISSN: 0025-1909. DOI: 10.1287/mnsc.2016.2503. URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2503 (visited on 11/28/2017).

Coppock, Alexander (2017). "Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment". In: *Statistics, Politics and Policy* 8.1, pp. 29–40. ISSN: 2194-6299. DOI: 10.1515/spp-2016-0005. URL: https://www.degruyter.com/view/j/spp.ahead-of-print/spp-2016-0005/spp-2016-0005.xml (visited on 12/05/2017).

Corstange, Daniel (2009). “Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT”. In: *Political Analysis* 17.1, pp. 45–63. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpn013. URL: https://www.cambridge.org/core/journals/political-analysis/article/sensitive-questions-truthful-answers-modeling-the-list-experiment-with-listit/D5914F542A81B1E4ADE0B53CDC17BD0B (visited on 12/02/2017).

Curran, Paul G (2016). “Methods for the detection of carelessly invalid responses in survey data”. In: *Journal of Experimental Social Psychology* 66, pp. 4–19.

Curran, Paul G and Kelsey A Hauser (2019). “I’m paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items”. In: *Journal of Research in Personality* 82, p. 103849.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian (2009). *Internet, mail, and mixed-mode surveys: The tailored design method, 3rd ed.* Internet, mail, and mixed-mode surveys: The tailored design method, 3rd ed. Hoboken, NJ, US: John Wiley & Sons Inc. ISBN: 978-0-471-69868-5.

Downes-Le Guin, Theo (2005). *Satisficing behavior in online panelists*. Chicago, IL.

Droitcour, Judith et al. (1991). “The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application”. en. In: *Measurement Errors in Surveys*. Ed. by Paul P. Biemer et al. John Wiley & Sons, Inc., pp. 185–210. ISBN: 978-1-118-15038-2. DOI: 10.1002/9781118150382.ch11.

Eady, Gregory (2016). “Replication Data for: The Statistical Analysis of Misreporting on Sensitive Survey Questions”. en. In: type: dataset. DOI: 10.7910/DVN/PZKBUX. URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PZKBUX (visited on 09/16/2018).

– (2017). “The statistical analysis of misreporting on sensitive survey questions”. In: *Political Analysis* 25.2, pp. 241–259.

Edwards, Allen Louis (1957). *The Social Desirability Variable in Personality Assessment and Research*. en. Google-Books-ID: 1jVIAAAAMAAJ. Dryden Press.

Fisher, Robert J. (1993). “Social Desirability Bias and the Validity of Indirect Questioning”. In: *Journal of Consumer Research* 20.2, pp. 303–315. ISSN: 0093-5301. URL: http://www.jstor.org/stable/2489277 (visited on 12/17/2017).

Frye, Timothy et al. (2016). “Is Putin’s popularity real?” In: *Post-Soviet Affairs* 33.1, pp. 1–15. ISSN: 1060-586X. DOI: 10.1080/1060586X.2016.1144334. URL: https://doi.org/10.1080/1060586X.2016.1144334 (visited on 12/07/2017).

Ghitza, Yair and Andrew Gelman (2020). “Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research”. In: *Political Analysis* 28.4, pp. 507–531. DOI: 10.1017/pan.2020.3.

Glynn, Adam N. (2010). “What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment”. In: *Unpublished manuscript*.

– (2013). “What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment”. In: *Public Opinion Quarterly* 77.S1, pp. 159–172. ISSN: 0033-362X. DOI: 10.1093/poq/nfs070. URL: https://academic.oup.com/poq/article/77/S1/159/1878470 (visited on 11/27/2017).

González-Ocantos, Ezequiel, Chad Kiewiet de Jonge, et al. (2012). “Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua”. en. In: *American Journal of Political Science* 56.1, pp. 202–217. ISSN: 1540-5907. DOI: 10.1111/j.1540-5907.2011.00540.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2011.00540.x/abstract (visited on 11/28/2017).

González-Ocantos, Ezequiel, Chad Kiewiet de Jonge, and David W. Nickerson (2015). “Legitimacy Buying: The Dynamics of Clientelism in the Face of Legitimacy Challenges”. en. In: *Comparative Political Studies* 48.9, pp. 1127–1158. ISSN: 0010-4140. DOI: 10.1177/0010414015574882. URL: https://doi.org/10.1177/0010414015574882 (visited on 11/28/2017).

Green, Donald P and Alan S Gerber (2006). “Can registration-based sampling improve the accuracy of midterm election forecasts?” In: *Public Opinion Quarterly* 70.2, pp. 197–223.

Groves, Robert M. and Lars Lyberg (2010). “Total Survey Error: Past, Present, and Future”. en. In: *Public Opinion Quarterly* 74.5, pp. 849–879. ISSN: 0033-362X. DOI: 10.1093/poq/nfq065. URL: https://academic.oup.com/poq/article/74/5/849/1817502 (visited on 06/14/2018).

Hauser, David J and Norbert Schwarz (2015). “It’sa trap! Instructional manipulation checks prompt systematic thinking on “tricky” tasks”. In: *Sage Open* 5.2, p. 2158244015584617.

Heerwig, Jennifer A. and Brian J. McCabe (2009). “Education and Social Desirability Bias: The Case of a Black Presidential Candidate”. en. In: *Social Science Quarterly* 90.3, pp. 674–686. ISSN: 1540-6237. DOI: 10.1111/j.1540-6237.2009.00637.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6237.2009.00637.x/abstract (visited on 12/05/2017).

Holbrook, Allyson L. and Jon A. Krosnick (2010). “Social desirability bias in voter turnout reports: Tests using the item count technique”. In: *Public Opinion Quarterly* 74.1, pp. 37–67. ISSN: 0033-362X. DOI: 10.1093/poq/nfp065. URL: https://academic.oup.com/poq/article/74/1/37/1841959 (visited on 12/03/2017).

Huang, Jason L. et al. (2012). “Detecting and Deterring Insufficient Effort Responding to Surveys”. en. In: *Journal of Business and Psychology* 27.1, pp. 99–114. ISSN: 0889-3268, 1573-353X. DOI: 10.1007/s10869-011-9231-8. URL:

https://link.springer.com/article/10.1007/s10869-011-9231-8 (visited on 12/17/2017).

Imai, Kosuke (2011). "Multivariate Regression Analysis for the Item Count Technique". In: *Journal of the American Statistical Association* 106.494, pp. 407–416. ISSN: 0162-1459. URL: http://www.jstor.org/stable/41416378 (visited on 11/27/2017).

Imai, Kosuke, Bethany Park, and Kenneth F. Greene (2015). "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models". In: *Political Analysis* 23.2, pp. 180–196. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpu017. URL: https://www.cambridge.org/core/journals/political-analysis/article/using-the-predicted-responses-from-list-experiments-as-explanatory-variables-in-regression-models/9696A7F03D046AC70D4F5C647D9AAACE (visited on 11/28/2017).

Imbens, Guido W. and Charles F. Manski (2004). "Confidence Intervals for Partially Identified Parameters". en. In: *Econometrica* 72.6, pp. 1845–1857. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2004.00555.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2004.00555.x/abstract (visited on 11/28/2017).

Johnson, John A. (2005). "Ascertaining the validity of individual protocols from Web-based personality inventories". en. In: *Journal of Research in Personality* 39.1, pp. 103–129. ISSN: 00926566. DOI: 10.1016/j.jrp.2004.09.009. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092656604000856 (visited on 06/03/2018).

Jones, Michael S., Lisa A. House, and Zhifeng Gao (2015). "Attribute Non-Attendance and Satisficing Behavior in Online Choice Experiments". en. In: *Proceedings in Food System Dynamics* 0.0, pp. 415–432. ISSN: 2194-511X. DOI: 10.18461/pfsd.2015.1534. URL: http://centmapress.ilb.uni-bonn.de/ojs/index.php/proceedings/article/view/1534 (visited on 12/17/2017).

Kane, James G., Stephen C. Craig, and Kenneth D. Wald (2004). "Religion and Presidential Politics in Florida: A List Experiment". en. In: *Social Science Quarterly* 85.2, pp. 281–293. ISSN: 1540-6237. DOI: 10.1111/j.0038-4941.2004.08502004.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.0038-4941.2004.08502004.x/abstract (visited on 12/03/2017).

Kapelner, Adam and Dana Chandler (2010). "Preventing Satisficing in Online Surveys: A "Kapcha" to Ensure Higher Quality Data". In: *Proceedings of CrowdConf 2010*. San Francisco, CA.

Kiewiet de Jonge, Chad P. (2015). "Who Lies About Electoral Gifts? Experimental Evidence from Latin America". In: *Public Opinion Quarterly* 79.3, pp. 710–739. ISSN: 0033-362X. DOI: 10.1093/poq/nfv024. URL: https://academic.oup.com/poq/article/79/3/710/1917596 (visited on 11/28/2017).

Kiewiet de Jonge, Chad P. and David W. Nickerson (2014). "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys". en. In: *Political Behavior* 36.3, pp. 659–682. ISSN: 0190-9320, 1573-6687. DOI: 10.1007/s11109-013-9249-x. URL: https://link.springer.com/article/10.1007/s11109-013-9249-x (visited on 11/28/2017).

Kim, Seo-young Silvia, Spencer Schneider, and R Michael Alvarez (2020). "Evaluating the Quality of Changes in Voter Registration Databases". In: *American Politics Research* 48.6, pp. 670–676.

King, Gary et al. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation". en. In: *American Political Science Review* 95.1, pp. 49–69. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/analyzing-incomplete-political-science-data-an-alternative-algorithm-for-multiple-imputation/9E712982CCE2DE79A574FE98488F212B (visited on 09/16/2018).

Köszegi, Botond (2006). "Ego Utility, Overconfidence, and Task Choice". In: *Journal of the European Economic Association* 4.4, pp. 673–707. ISSN: 15424766. URL: https://clsproxy.library.caltech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=21615516&site=ehost-live&scope=site (visited on 09/03/2018).

Kramon, Eric and Keith R. Weghorst (2012). "Measuring sensitive attitudes in developing countries: lessons from implementing the list experiment". In: *Newsletter of the APSA Experimental Section* 3.2, pp. 14–24.

Krosnick, Jon A (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". In: *Applied cognitive psychology* 5.3, pp. 213–236.

Kuhn, Patrick M and Nick Vivyan (2021). "The misreporting trade-off between list experiments and direct questions in practice: Partition validation evidence from two countries". In: *Political Analysis*, pp. 1–22. DOI: doi:10.1017/pan.2021.10.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens (1997). "Racial Attitudes and the "New South"". In: *The Journal of Politics* 59.2, pp. 323–349. ISSN: 0022-3816. DOI: 10.2307/2998167. URL: http://www.jstor.org/stable/2998167 (visited on 12/02/2017).

Kung, Franki YH, Navio Kwok, and Douglas J Brown (2018). "Are attention check questions a threat to scale validity?" In: *Applied Psychology* 67.2, pp. 264–283.

Lax, Jeffrey R., Justin H. Phillips, and Alissa F. Stollwerk (2016). "Are Survey Respondents Lying about Their Support for Same-Sex Marriage? Lessons from a List Experiment". In: *Public Opinion Quarterly* 80.2, pp. 510–533. ISSN: 0033-362X. DOI: 10.1093/poq/nfv056. URL: https://academic.oup.com/poq/article/80/2/510/2588811 (visited on 11/28/2017).

Little, Roderick J. A. (1992). “Regression With Missing X’s: A Review”. In: *Journal of the American Statistical Association* 87.420, pp. 1227–1237. ISSN: 0162-1459. DOI: 10.2307/2290664. URL: https://www.jstor.org/stable/2290664 (visited on 09/16/2018).

Maccoby, Eleanor E., Nathan Maccoby, and Gardner Lindzey (1954). “The interview: A tool of social science”. In: *Handbook of Social Psychology*. Vol. 1. Addison-Wesle, pp. 449–487.

Malesky, Edmund J., Dimitar D. Gueorguiev, and Nathan M. Jensen (2015). “Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment”. en. In: *American Journal of Political Science* 59.2, pp. 419–439. ISSN: 1540-5907. DOI: 10.1111/ajps.12126. URL: http://onlinelibrary.wiley.com/doi/10.1111/ajps.12126/abstract (visited on 11/28/2017).

Maniaci, Michael R and Ronald D Rogge (2014). “Caring about carelessness: Participant inattention and its effects on research”. In: *Journal of Research in Personality* 48, pp. 61–83.

Meade, Adam W and S Bartholomew Craig (2012). “Identifying careless responses in survey data.” In: *Psychological methods* 17.3, p. 437.

Meng, Tianguang, Jennifer Pan, and Ping Yang (2014). “Conditional Receptivity to Citizen Participation: Evidence From a Survey Experiment in China”. en. In: *Comparative Political Studies* 50.4, pp. 399–433. ISSN: 0010-4140. DOI: 10.1177/0010414014556212. URL: https://doi.org/10.1177/0010414014556212 (visited on 11/28/2017).

Miller, Jeff (2006). “Research reveals alarming incidence of ‘undesirable’ online panelists”. In: *Research Conference Report, September-October 2006, RFL Communications. Synthesized from presentation at Research Industry Summit: Improving Respondent Cooperation, Chicago, IL, USA Sept 28*. Vol. 29, p. 2006.

Miller, Judith Droitcour (1984). “A new survey technique for studying deviant behavior”. English. OCLC: 12768382. PhD thesis.

Oppenheimer, Daniel M, Tom Meyvis, and Nicolas Davidenko (2009). “Instructional manipulation checks: Detecting satisficing to increase statistical power”. In: *Journal of experimental social psychology* 45.4, pp. 867–872.

Ostwald, Kai and Guillem Riambau (2018). “Placebo Statements in List Experiments”. In: *Unpublished manuscript*.

Pepinsky, Thomas B (2018). “A note on listwise deletion versus multiple imputation”. In: *Political Analysis* 26.4, pp. 480–488.

Read, Blair, Lukas Wolters, and Adam J. Berinsky (2021). “Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys”. In: *Political Analysis*. DOI: 10.1017/pan.2021.32.

Redlawsk, David P., Caroline J. Tolbert, and William Franko (2010). “Voters, Emotions, and Race in 2008: Obama as the First Black President”. en. In: *Political Research Quarterly* 63.4, pp. 875–889. ISSN: 1065-9129. DOI: 10.1177/1065912910373554. URL: https://doi.org/10.1177/1065912910373554 (visited on 12/03/2017).

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro (2016). “An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions”. en. In: *American Journal of Political Science* 60.3, pp. 783–802. ISSN: 1540-5907. DOI: 10.1111/ajps.12205. URL: http://onlinelibrary.wiley.com/doi/10.1111/ajps.12205/abstract (visited on 11/28/2017).

Simon, Herbert A (1956). “Rational choice and the structure of the environment.” In: *Psychological review* 63.2, p. 129.

Stoye, Jörg (2009). “More on Confidence Intervals for Partially Identified Parameters”. en. In: *Econometrica* 77.4, pp. 1299–1315. ISSN: 1468-0262. DOI: 10.3982/ECTA7347. URL: http://onlinelibrary.wiley.com/doi/10.3982/ECTA7347/abstract (visited on 11/28/2017).

Streb, Matthew J. et al. (2008). “Social Desirability Effects and Support for a Female American President”. In: *Public Opinion Quarterly* 72.1, pp. 76–89. ISSN: 0033-362X. DOI: 10.1093/poq/nfm035. URL: https://academic.oup.com/poq/article/72/1/76/1816464 (visited on 12/04/2017).

Thomas, Kyle A. and Scott Clifford (2017). “Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments”. In: *Computers in Human Behavior* 77, pp. 184–197. DOI: https://doi.org/10.1016/j.chb.2017.08.038.

Tourangeau, Roger and Tom W Smith (1996). “Asking sensitive questions: The impact of data collection mode, question format, and question context”. In: *Public opinion quarterly* 60.2, pp. 275–304.

Vannette, David (2017). *Using Attention Checks in Your Surveys May Harm Data Quality*. en-US. URL: https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/ (visited on 06/14/2018).

Ward, M.K. and Samuel B. Pond (2015). “Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys”. en. In: *Computers in Human Behavior* 48, pp. 554–568. ISSN: 07475632. DOI: 10.1016/j.chb.2015.01.070. URL: http://linkinghub.elsevier.com/retrieve/pii/S0747563215000989 (visited on 06/04/2018).

Warner, Stanley L. (1965). “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias”. In: *Journal of the American Statistical Association* 60.309, pp. 63–69. ISSN: 0162-1459. DOI: 10.2307/2283137. URL: http://www.jstor.org/stable/2283137 (visited on 12/17/2017).

Zagorsky, Jay L. and Patricia Rhoton (2008). “The Effects of Promised Monetary Incentives on Attrition in a Long-Term Panel Survey”. In: *Public Opinion Quarterly* 72.3, pp. 502–513. ISSN: 0033-362X. DOI: `10.1093/poq/nfn025`. URL: `https://academic.oup.com/poq/article/72/3/502/1835648` (visited on 12/17/2017).

Appendix A

APPENDIX TO CHAPTER II

Analytical Solution to Linear System (2)

In this section, I detail the steps to obtain the crude bound from linear system (2):

$$\begin{aligned}
&\max/\min_{(p_{kL},p_{kN},p_{kT})_{k=0}^{J}} \textstyle\sum_{k=0}^{J} p_{kL} + p_{kT} \\
&\text{s.t. } c_0 = p_{0N} + p_{0L} + p_{0T},\ \ldots,\ c_J = p_{JN} + p_{JL} + p_{JT} \\
&\quad t_0 = p_{0N} + p_{0L},\ \ldots,\ t_J = p_{JN} + p_{JL} + p_{J-1,T},\ t_{J+1} = p_{JT} \\
&\quad p_{kL}/(p_{kL} + p_{kT}) \leq (p_{JN} + p_{JL})/(p_{JN} + p_{JL} + p_{JT}),\ \forall k = 0, \ldots, J-1.
\end{aligned}$$

Step 1: Identify the sum of non-supporters and lying supporters $p_{kN} + p_{kL}$ and truth-telling supporters p_{kT}, $k = 1, ..., J$, from the distribution of responses under control $\{c_k\}_{k=0}^{J}$ and treated $\{t_k\}_{k=0}^{J+1}$.

$c_0 = p_{0N} + p_{0L} + p_{0T}$	$t_0 = p_{0N} + p_{0L}$
$c_1 = p_{1N} + p_{1L} + p_{1T}$	$t_1 = p_{1N} + p_{1L} + p_{0T}$
$c_2 = p_{2N} + p_{2L} + p_{2T}$	$t_2 = p_{2N} + p_{2L} + p_{1T}$
$c_3 = p_{3N} + p_{3L} + p_{3T}$	$t_3 = p_{3N} + p_{3L} + p_{2T}$
$c_4 = p_{4N} + p_{4L} + p_{4T}$	$t_4 = p_{4N} + p_{4L} + p_{3T}$
	$t_5 = p_{4T}$

Table A.1: Relationship between latent attitudes and observable responses ($J = 4$)

In particular, I obtain

$$p_{kT} = \sum_{i=k+1}^{J+1} t_i - \sum_{j=k+1}^{J} c_j, \quad p_{kN} + p_{kL} = \sum_{j=k}^{J} c_j - \sum_{i=k+1}^{J+1} t_i. \tag{A.1}$$

Step 2: Calculate the maximal liar ratio used to construct the crude bound under the relaxed liars assumption:

$$\lambda \equiv \frac{p_{JN} + p_{JL}}{p_{JN} + p_{JL} + p_{JT}} = \frac{c_J - t_{J+1}}{c_J}. \tag{A.2}$$

Step 3: Calculate the maximal proportion of liars for respondents answering affirmatively to fewer than J control items:[1]

$$p_{kL}/(p_{kL} + p_{kT}) \leq \lambda, \tag{A.3}$$

[1] For list experiments with a modest sample size, $\hat{p}_{kT}$ may be negative even if p_{kT} is positive in the true data generating process. For cases where $\hat{p}_{kT} < 0$, I set $\hat{p}_{kL} = 0$.

$$\Rightarrow p_{kL} \leq \frac{\lambda}{1-\lambda} p_{kT}. \tag{A.4}$$

Step 4: Lower bound and upper bound are given by:

$$\sum_{k=0}^{J} p_{kT} = \sum_{k=0}^{J} (\sum_{i=k+1}^{J+1} t_i - \sum_{j=k+1}^{J} c_j) = \sum_{i=0}^{J+1} it_i - \sum_{j=0}^{J} jc_j, \tag{A.5}$$

$$\sum_{k=0}^{J} p_{kT} + p_{kL} \leq \sum_{k=0}^{J} p_{kT} + \sum_{k=0}^{J} \min\{\frac{\lambda}{1-\lambda} p_{kT}, p_{kN} + p_{kL}\} \tag{A.6}$$

$$= (\sum_{i=0}^{J+1} it_i - \sum_{j=0}^{J} jc_j) + \sum_{k=0}^{J} \min\{\frac{\lambda}{1-\lambda} (\sum_{i=k+1}^{J+1} t_i - \sum_{j=k+1}^{J} c_j), \sum_{j=k}^{J} c_j - \sum_{i=k+1}^{J+1} t_i\}, \tag{A.7}$$

where the lower bound is the standard difference in means estimate, and the upper bound is weakly smaller than $1/(1-\lambda)$ multiplied by the lower bound.

Tables for Illustrative Example

Table A.2: Organizations included in the list experiment

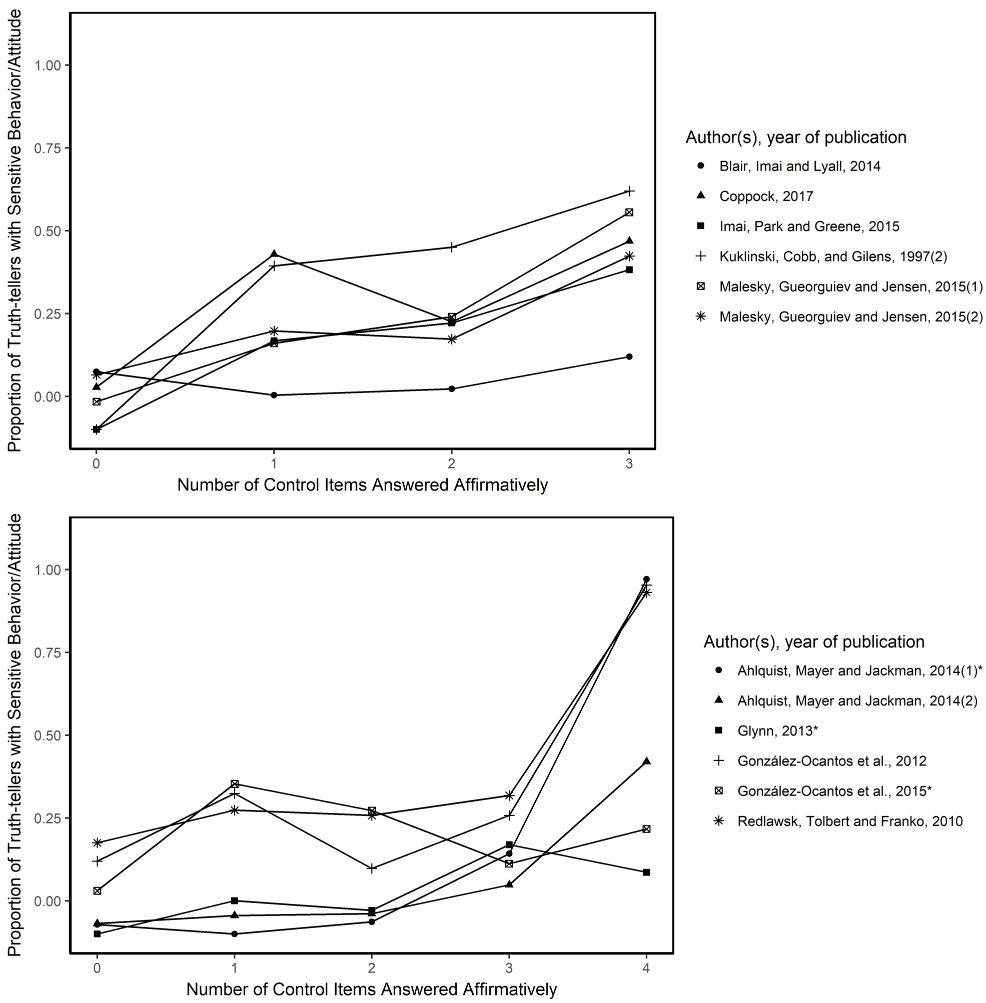
Item Type	Name (Short Description)
Control	*Californians for Disability Rights* (organization advocating for people with disabilities) *California National Organization for Women* (organization advocating for women's equality and empowerment) *American Family Association* (organization advocating for pro-family values) *American Red Cross* (humanitarian organization)
X-treatment	*Organization X* (organization advocating for immigration reduction and measures against undocumented immigration)
Y-treatment	*Organization Y* (citizen border patrol group combating undocumented immigration)

Table A.3: Distribution of responses under control, X-treatment and Y-treatment

	0	1	2	3	4	5
Control	.13	.16	.22	.25	.24	
X-treatment	.12	.11	.21	.24	.15	.17
Y-treatment	.12	.15	.20	.27	.12	.14

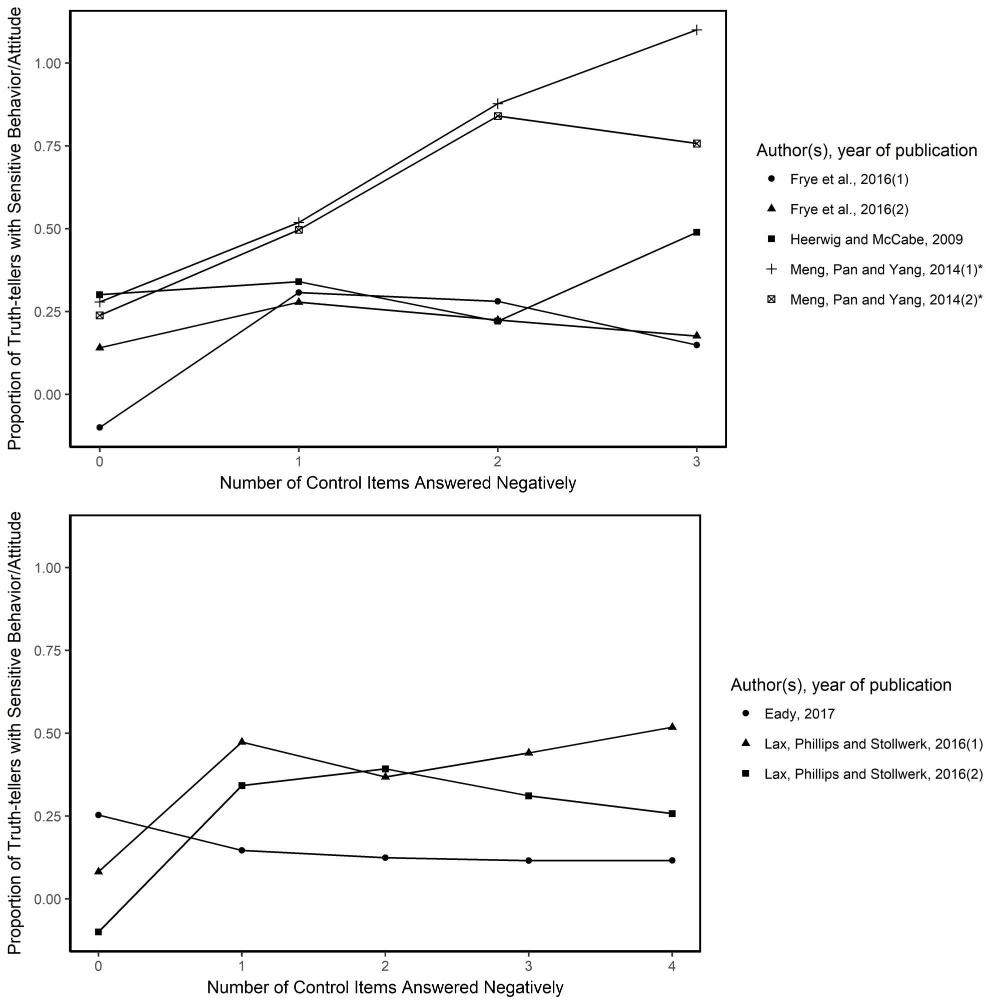
Figures for List Experiments in Published Studies

Figure A.1: Proportion of respondents with sensitive behavior/attitude (the sensitive response is affirmative) conditional on the number of control items answered affirmatively



Note: The figure presents the proportion of truth-tellers with sensitive behavior/attitude conditional on the number of control items answered affirmatively for list experiments with affirmative sensitive response (top panel: $J = 3$, bottom panel: $J = 4$). List experiments with fewer than 50 respondents choosing the maximal number of items are marked with stars. Proportions smaller than -0.1 are trimmed for graphical presentation.

Figure A.2: Proportion of respondents with sensitive behavior/attitude (the sensitive response is negative) conditional on the number of control items answered negatively



Note: The figure presents the proportion of truth-tellers with sensitive behavior/attitude conditional on the number of control items answered negatively for list experiments with negative sensitive response (top panel: $J = 3$, bottom panel: $J = 4$). List experiments with fewer than 50 respondents choosing the minimal number of items are marked with stars. Proportions smaller than -0.1 or larger than 1.1 are trimmed for graphical presentation.

Additional Simulations

High vs. Low Correlation between Sensitive Item and Control Items

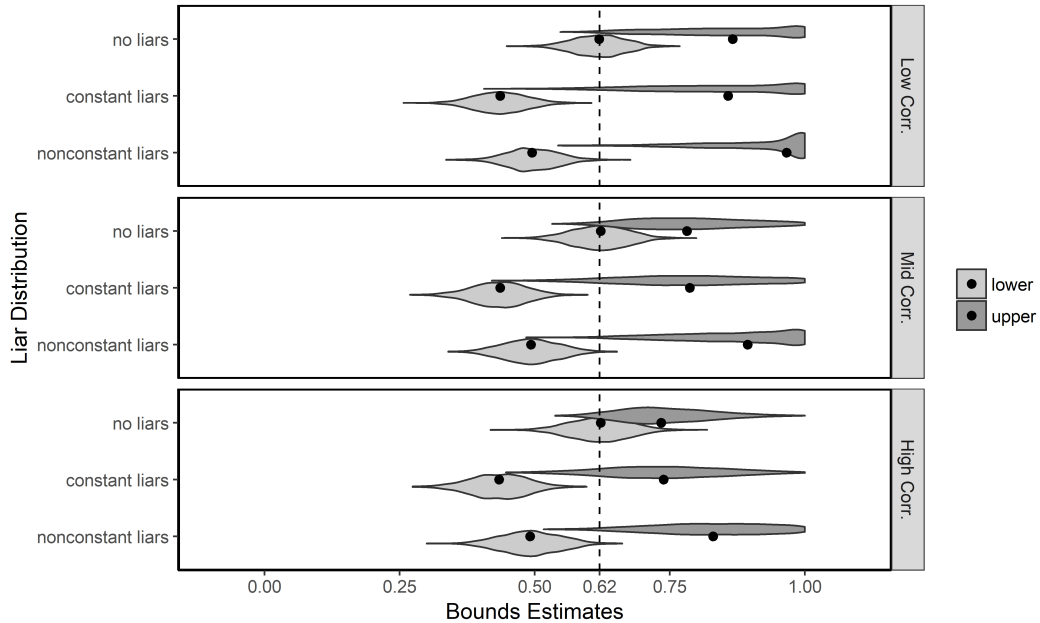


Figure A.3: Bound Estimates for Simulated List Experiments, II

Note: The figure presents the densities of lower bound and upper bound estimates for 2000 simulated datasets for each of the 9 types of list experiments. The high, medium, and low correlation lists are otherwise identical to the correlated lists except that the added pairwise correlation between items are 0.15, 0.1, and 0.05, respectively (instead of 0.1). Lower bound densities are shown in lighter grey and upper bound densities are shown in darker grey. Black dots are the median lower bound and upper bound estimates.

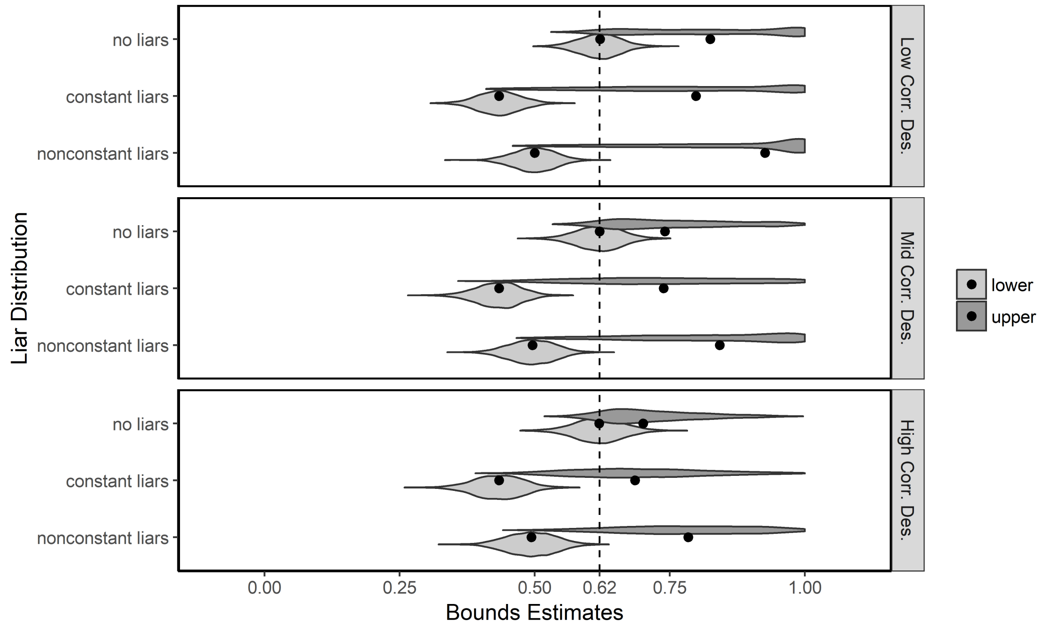


Figure A.4: Bound Estimates for Simulated List Experiments, III

Note: The figure presents the densities of lower bound and upper bound estimates for 2000 simulated datasets for each of the 9 types of list experiments. The high, medium, and low correlation lists are otherwise identical to the correlated design lists except that the added pairwise correlation between items are 0.15, 0.1, and 0.05, respectively. Lower bound densities are shown in lighter grey and upper bound densities are shown in darker grey. Black dots are the median lower bound and upper bound estimates.

High vs. Low Prevalence of the Sensitive Item

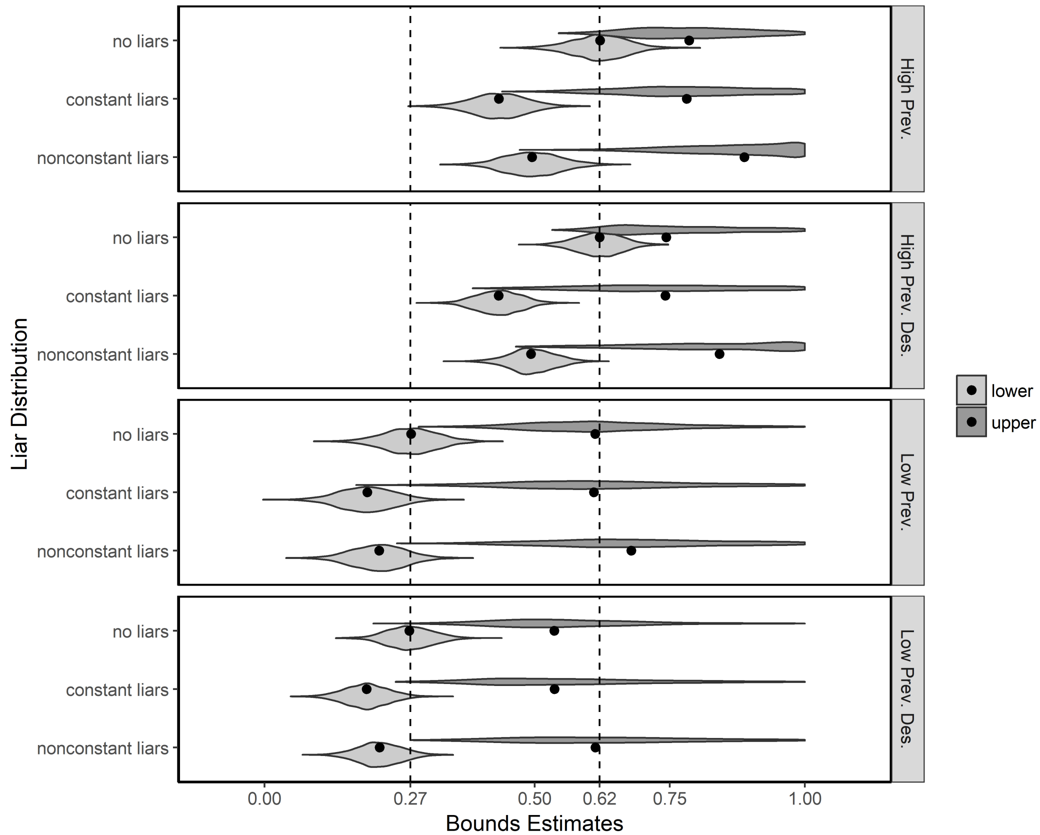


Figure A.5: Bound Estimates for Simulated List Experiments, IV

Note: The figure presents the densities of lower bound and upper bound estimates for 2000 simulated datasets for each of the 12 types of list experiments. The high and low prevalence correlated lists/correlated design lists are otherwise identical to the correlated lists/correlated design lists except that the prevalence of the sensitive item is 0.62 and 0.27, respectively. Lower bound densities are shown in lighter grey and upper bound densities are shown in darker grey. Black dots are the median lower bound and upper bound estimates.

Affirmative vs. Negative Sensitive Responses

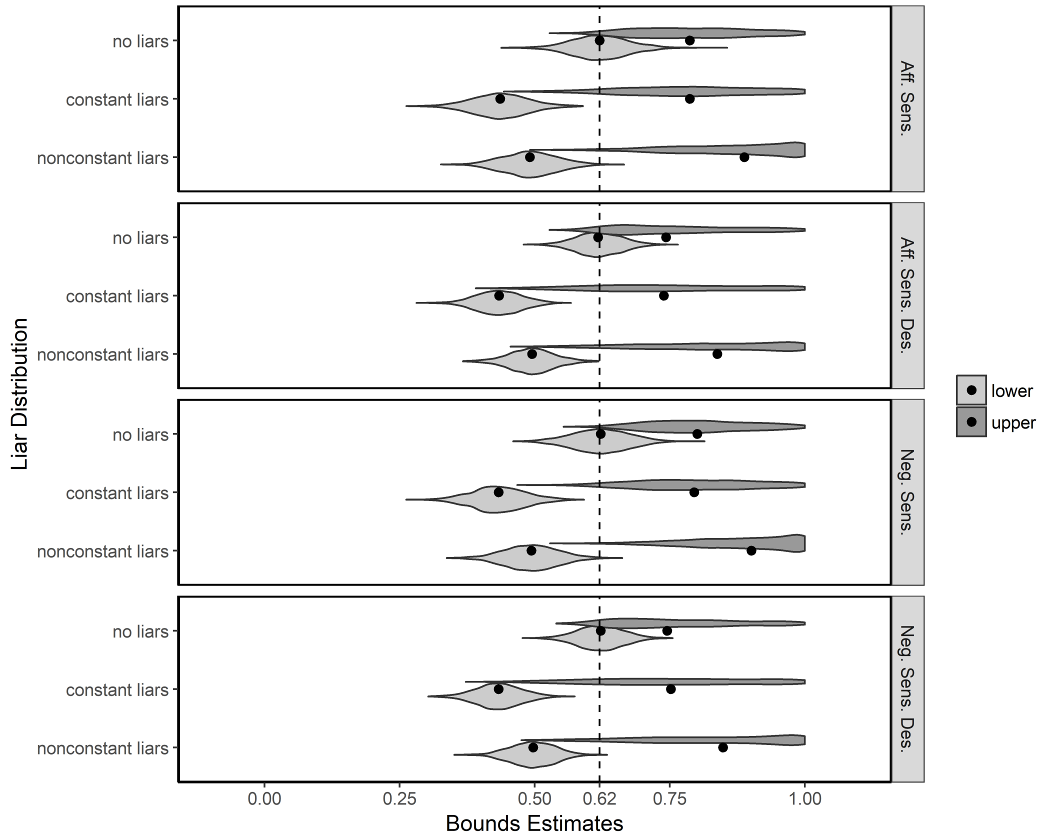


Figure A.6: Bound Estimates for Simulated List Experiments, V

Note: The figure presents the densities of lower bound and upper bound estimates for 2000 simulated datasets for each of the 12 types of list experiments. The affirmative and negative sensitive correlated lists/correlated design lists are otherwise identical to the correlated lists/correlated design lists except that the sensitive response is affirmative and negative, respectively. Lower bound densities are shown in lighter grey and upper bound densities are shown in darker grey. Black dots are the median lower bound and upper bound estimates.

List Experiments with Negative Sensitive Responses

I can readily adapt my method to list experiments where a negative answer to the sensitive item is considered sensitive. As in the text, I denote the population fraction of respondents who answer affirmatively to k control items and (1) without the sensitive behavior or attitude by p_{kN}; (2) with the sensitive behavior or attitude, but would *not* give the truthful answer for it by p_{kL}; (3) with the sensitive behavior or attitude, and give the truthful answer for it by p_{kT}. The difference from the case of affirmative sensitive responses is that the truthful answer is negative.

Similar to the case of affirmative sensitive responses, I can establish the relationship between the proportions of different types of respondents and the distribution of answers under control and treatment, shown in table A.4. It follows from a close inspection of the equations that I can identify $p_{kN}+p_{kL}$ and p_{kT} for each $k = 1, ..., J$.

$c_0 = p_{0N} + p_{0L} + p_{0T}$	$t_0 = p_{0T}$
$c_1 = p_{1N} + p_{1L} + p_{1T}$	$t_1 = p_{0N} + p_{0L} + p_{1T}$
$c_2 = p_{2N} + p_{2L} + p_{2T}$	$t_2 = p_{1N} + p_{1L} + p_{2T}$
$c_3 = p_{3N} + p_{3L} + p_{3T}$	$t_3 = p_{2N} + p_{2L} + p_{3T}$
$c_4 = p_{4N} + p_{4L} + p_{4T}$	$t_4 = p_{3N} + p_{3L} + p_{4T}$
	$t_5 = p_{4N} + p_{4L}$

Table A.4: Relationship between latent attitudes and observable responses ($J = 4$) for list experiments with negative sensitive responses

For list experiments with negative sensitive responses, truth-telling is fully revealing about the sensitive item for respondents whose answer is negative to all control items (floor effects). I consider the relaxed liars assumption that states among all respondents with the sensitive behavior or attitude, the ones who respond negatively to all control items have the strongest incentive to lie:

$$\frac{p_{kL}}{p_{kL} + p_{kT}} \leq \frac{p_{0L}}{p_{0L} + p_{0T}}, \ \forall k = 1, \ldots, J. \tag{A.8}$$

Similar to the case of affirmative sensitive responses, a crude upper/lower bound for the level of support for the sensitive item is given by the solution to the following linear system:

$$\begin{aligned}
&\max/\min_{(p_{kL}, p_{kN}, p_{kT})_{k=0}^{J}} \textstyle\sum_{k=0}^{J} p_{kL} + p_{kT} \\
&\text{s.t. } c_0 = p_{0N} + p_{0L} + p_{0T}, \ \ldots, \ c_J = p_{JN} + p_{JL} + p_{JT} \\
&\qquad t_0 = p_{0T}, \ \ldots, \ t_J = p_{J-1,N} + p_{J-1,L} + p_{J,T}, \ t_{J+1} = p_{JN} + p_{JL} \\
&\qquad p_{kL}/(p_{kL} + p_{kT}) \leq (p_{0N} + p_{0L})/(p_{0N} + p_{0L} + p_{0T}), \ \forall k = 1, \ldots, J.
\end{aligned} \tag{A.9}$$

The confidence set for the interval estimate can be constructed in the same way as before.

For the mappings in Table A.4 and the relaxed liars assumption (A.8), they are mirror-symmetric to the case of affirmative sensitive responses (Table 3 and Inequality (1)). For ease of exposition, in the text I consider proportions of different types of respondents conditional on the number of control items answered negatively (instead of affirmatively). For example, answering negatively to all control items is equivalent to answering affirmatively to zero control items.

No Design Effect

While relaxing the no liars assumption, I maintain the assumption of no design effect. No design effect is more likely to be satisfied if respondents consider the items on the list one by one and "do not evaluate items on the list relative to one another" (Imai, 2011, p. 409). While allowing the possibility of liars in my analysis, no design effect still rules out inter-item behavior like concealing preference for the sensitive item by lying about nonsensitive items (e.g., by choosing zero item). However, there is no consensus on how no design effect is likely to be violated in list experiments, which makes it difficult, if at all possible, to develop techniques robust to such violations.

Meanwhile, some researchers (Holbrook and Jon A. Krosnick, 2010; Ahlquist, Mayer, and Jackman, 2014; C. P. Kiewiet de Jonge and Nickerson, 2014; Frye et al., 2016) use placebo list experiments to detect violations of no design effects, where they replace the sensitive item with a nonsensitive item with prevalence either known or estimable. Placebo experiments allow, to some extent, a comparison of the average latent response to control items under control and treatment. If there is no significant difference, researchers may have more confidence in no design effect and focus on potential violations of no liars.

Summary of List Experiments in Section 3.2

Table A.5: Summary of List Experiments in Section 3.2

	Sensitive behavior or attitude	N_C	Yes/No	N	Mode
Kuklinski et al. (1997)	A black family moving in next door	3	Yes	1213	telephone
	Black leaders asking for affirmative action	3	Yes	1171	telephone
Heerwig and McCabe (2009)	Supporting a black presidential candidate	3	No	1044	online
Redlawsk et al. (2010)	First black president	4	Yes	1395	telephone
Gonzalez-Ocantos et al. (2012)	Vote buying in Nicaragua	4	Yes	995	face-to-face
Glynn (2013)	Black person becoming president	4	Yes	1762	online
Ahlquist et al. (2014)	Voter impersonation (wave 1)	4	Yes	995	online
	Voter impersonation (wave 2)	4	Yes	3000	online
Blair et al. (2014)	Support for ISAF among Pashtun men	3	Yes	1836	face-to-face
Meng et al. (2014)	Officials' receptivity to formal participation in China	3	No	883	private
	Officials' receptivity to Internet participation in China	3	No	868	private
Gonzalez-Ocantos et al. (2015)	Vote buying in Honduras	4	Yes	993	face-to-face
Imai et al. (2015)	Vote buying in Mexico	3	Yes	1120	face-to-face
Malesky et al. (2015)	Bribery in Vietnam (domestic firms)	3	Yes	16236	NA
	Bribery in Vietnam (foreign-invested enterprises)	3	Yes	3570	NA
Frye et al. (2016)	Support for Putin (historical list, March wave)	3	No	1599	telephone
	Support for Putin (contemporary list, March wave)	3	No	1598	telephone
Lax et al. (2016)	Support for same-sex marriage	4	No	1878	online
	Support for employment non-discrimination laws	4	No	1187	online
Rosenfeld et al. (2016)	Vote for an anti-abortion referendum (MS 2011)	4	Yes	1319	telephone
Coppock (2017)	Support for Trump	3	Yes	5290	online
Eady (2017)	Women's competence in politics	4	No	22372	online

Note: The table summarizes basic information about list experiments in published studies: the sensitive behaviors or attitudes measured (column 2), the number of control items N_C (column 3), whether an affirmative ('yes') or negative ('no') response to the sensitive item is considered sensitive (column 4), the sample size N (column 5), and the mode of the survey (column 6).

Appendix B

APPENDIX TO CHAPTER III

Figure B.1: Sample e-mail Invite

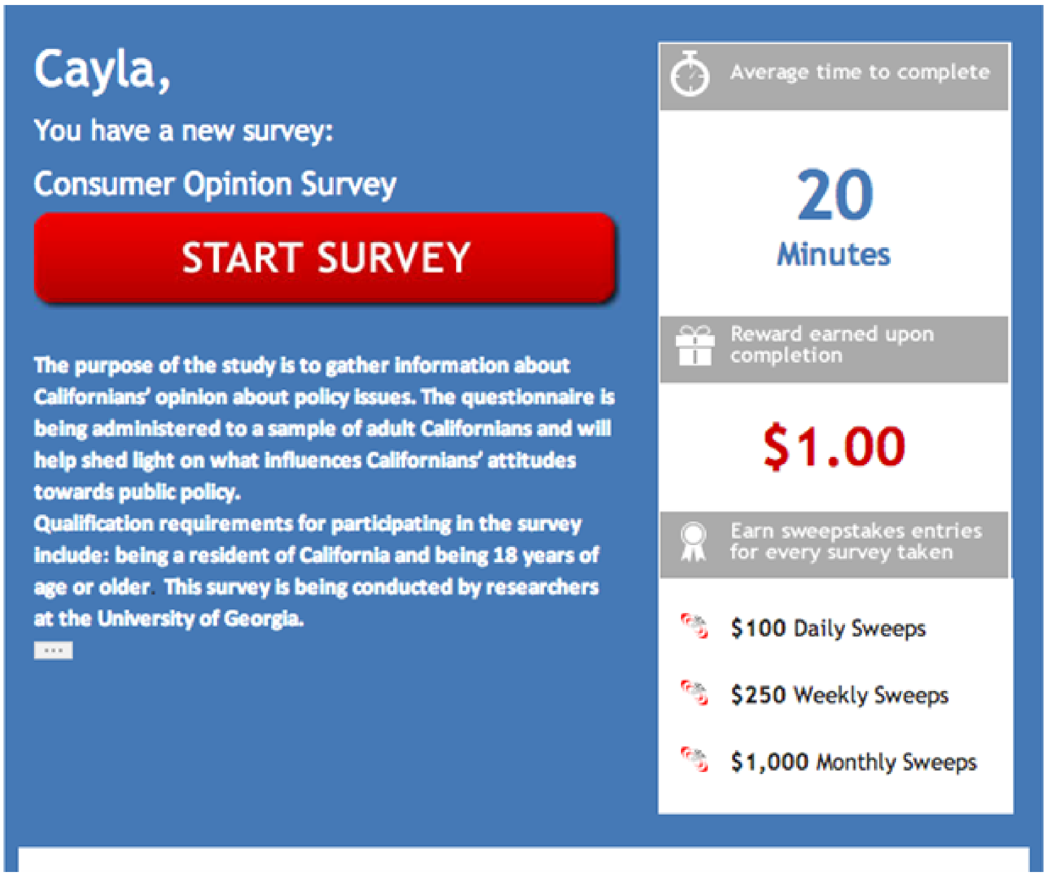
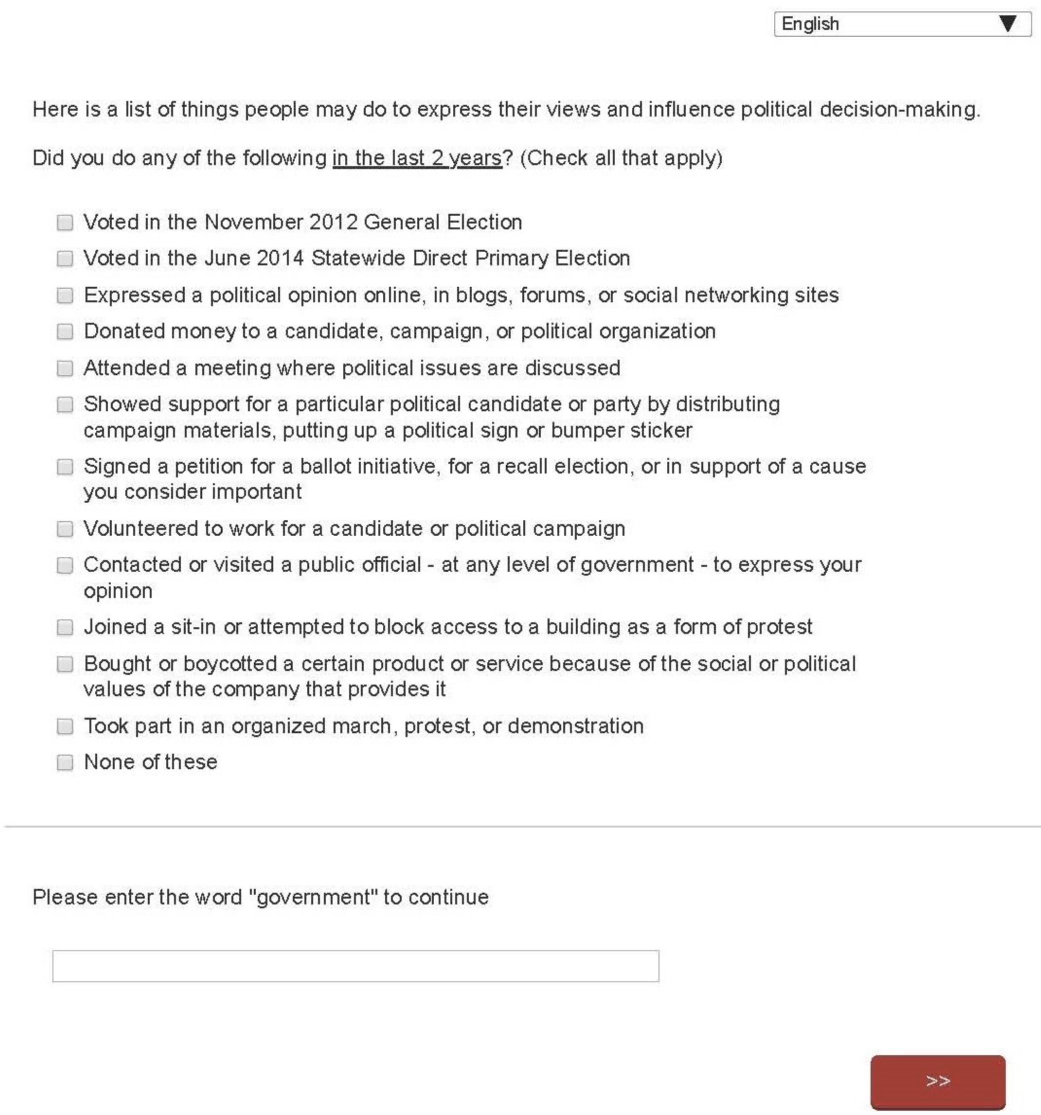


Figure B.2: Screenshot of First Attention Check (TQ 1, desktop version)



English

Here is a list of things people may do to express their views and influence political decision-making.

Did you do any of the following in the last 2 years? (Check all that apply)

- Voted in the November 2012 General Election
- Voted in the June 2014 Statewide Direct Primary Election
- Expressed a political opinion online, in blogs, forums, or social networking sites
- Donated money to a candidate, campaign, or political organization
- Attended a meeting where political issues are discussed
- Showed support for a particular political candidate or party by distributing campaign materials, putting up a political sign or bumper sticker
- Signed a petition for a ballot initiative, for a recall election, or in support of a cause you consider important
- Volunteered to work for a candidate or political campaign
- Contacted or visited a public official - at any level of government - to express your opinion
- Joined a sit-in or attempted to block access to a building as a form of protest
- Bought or boycotted a certain product or service because of the social or political values of the company that provides it
- Took part in an organized march, protest, or demonstration
- None of these

Please enter the word "government" to continue

>>

Figure B.3: Screenshot of Second Attention Check (TQ 2, desktop version)

English

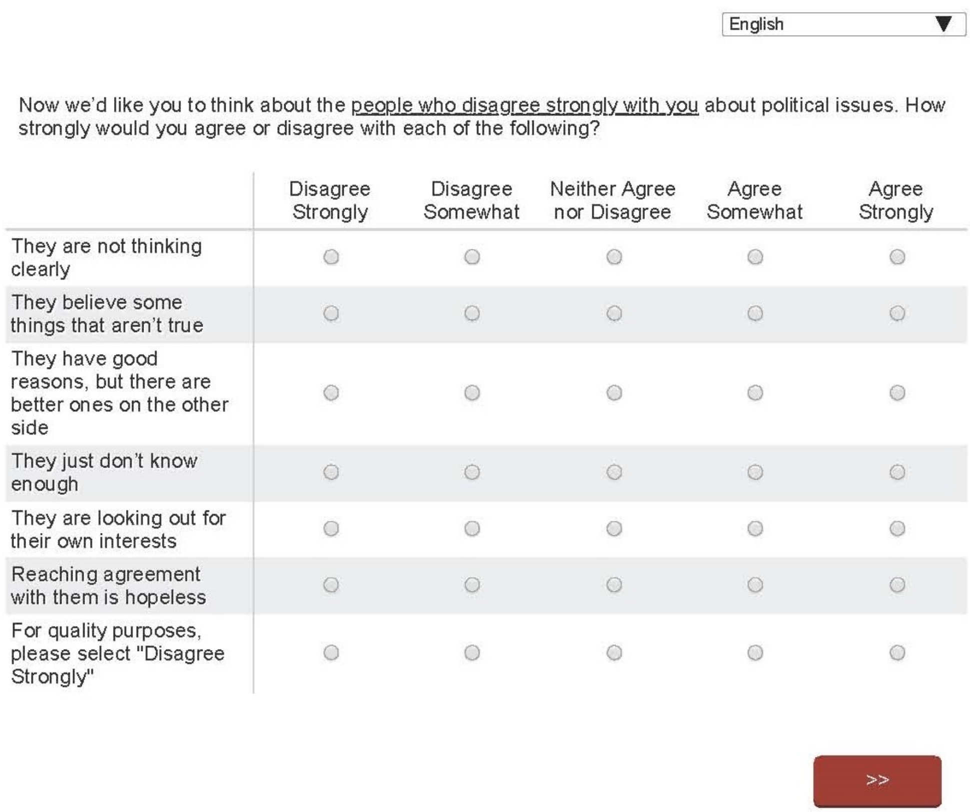
The federal government in Washington D.C. considered many important policy issues in recent years.

For each of the following policy issues tell us whether you support or oppose the regulation or legislation.

	Oppose	Support	I'm indifferent	I don't know
Implementing stricter carbon emission limits				
Providing a path to legal status and citizenship for undocumented immigrants				
Limiting National Security Agency's (NSA) collection of domestic phone records				
Repealing of *Don't Ask, Don't Tell*, allowing gays to serve openly in the armed services				
Restricting the sale of semi-automatic and automatic firearms, handguns, and high-capacity ammunition clips				
For quality purposes, please select "I'm indifferent"				
Health Care Reform, officially called the Patient Protection and Affordable Care Act				

>>

Figure B.4: Screenshot of Third Attention Check (TQ 3, desktop version)



English ▼

Now we'd like you to think about the people who disagree strongly with you about political issues. How strongly would you agree or disagree with each of the following?

	Disagree Strongly	Disagree Somewhat	Neither Agree nor Disagree	Agree Somewhat	Agree Strongly
They are not thinking clearly	○	○	○	○	○
They believe some things that aren't true	○	○	○	○	○
They have good reasons, but there are better ones on the other side	○	○	○	○	○
They just don't know enough	○	○	○	○	○
They are looking out for their own interests	○	○	○	○	○
Reaching agreement with them is hopeless	○	○	○	○	○
For quality purposes, please select "Disagree Strongly"	○	○	○	○	○

>>

Table B.1: How many fail?

	N	%
Pass	1,750	64
Fail	975	36
Fail TQ 1	575	21
Fail TQ 2	229	8
Fail TQ 3	171	6

Note: The table shows the number and percentage of respondents who passed and failed trap questions (TQs).

Table B.2: Who fails?

	All TQs		TQ 1 (Group I)		TQ 2 (Group II)		TQ 3 (Group III)	
Characteristic	Pass all	Fail any	Pass	Fail	Pass	Fail	Pass	Fail
Gender								
Female	64.1	35.9	78.5	21.5	89.8	10.2	90.9	9.1
Male	64.5	35.5	79.3	20.7	89	11	91.4	8.6
Education								
DNG HS	**53.8**	**46.2**	83.8	16.2	**82.1**	**17.9**	**78.2**	**21.8**
High School	**53.9**	**46.1**	75.8	24.2	**82.5**	**17.5**	**86.2**	**13.8**
Some College	**65.4**	**34.6**	78.2	21.8	**91.3**	**8.7**	**91.6**	**8.4**
College	**69.4**	**30.6**	81.6	18.4	**91.7**	**8.3**	**92.8**	**7.2**
Postgraduate	**69.8**	**30.2**	79.2	20.8	**91.7**	**8.3**	**96.2**	**3.8**
Age								
18 to 24	**46.8**	**53.2**	**76.9**	**23.1**	**74.7**	**25.3**	**81.5**	**18.5**
25 to 35	**56.6**	**43.4**	**74.2**	**25.8**	**87**	**13**	**87.7**	**12.3**
36 to 50	**65.2**	**34.8**	**79.2**	**20.8**	**89**	**11**	**92.5**	**7.5**
51 to 65	**74.2**	**25.8**	**81.5**	**18.5**	**96.4**	**3.6**	**94.4**	**5.6**
Older than 65	**79.5**	**20.5**	**83.7**	**16.3**	**97.5**	**2.5**	**97.4**	**2.6**
Region								
Bay Area	**66.2**	**33.8**	77.2	22.8	**92.6**	**7.4**	92.6	7.4
Central Valley	**64.3**	**35.7**	82.4	17.6	**84.6**	**15.4**	92.2	7.8
Cent./South. Farm	**63.4**	**36.6**	79.8	20.2	**87.7**	**12.3**	90.5	9.5
Nor. and Mount.	**59.4**	**40.6**	75.5	24.5	**87.2**	**12.8**	90.2	9.8
SoCal (excl. L.A.)	**68.2**	**31.8**	80.4	19.6	**92.8**	**7.2**	91.4	8.6
SoCal (L.A.)	**60.9**	**39.1**	77.6	22.4	**87.1**	**12.9**	90.2	9.8
N	1,750	972	2,148	574	1,921	227	1,750	171

Note: The table shows the percentage of respondents, within each demographic group, who passed or failed trap questions. Bold numbers indicate a statistically significant relationship (at a 95% confidence level) between the demographic attribute and the failure rate based on a Chi-squared test. "DNG HS" stands for "did not graduate from high school."

Table B.3: How do failers behave?

	All TQs		TQ 1 (Group I)		TQ 2 (Group II)		TQ 3 (Group III)	
Behavior	Pass all	Fail any	Pass	Fail	Pass	Fail	Pass	Fail
Avg. response time (s)	**35.8**	**22.2**	**31.9**	**27.2**	**33.9**	**15.6**	**35.8**	**14.2**
Non-attitudes rate (%)	**18.6**	**23.7**	20.2	21.3	**19.3**	**27.3**	**18.6**	**26.9**
Consistency of pref.								
Incomplete (%)	**9.1**	**18.0**	**11.6**	**14.6**	10.4	22.0	**9.1**	**24.0**
Intransitive (%)	**25.0**	**37.9**	**27.5**	**36.1**	**26.3**	**39.0**	**25.0**	**43.1**
N	1,750	972	2,148	574	1,921	227	1,750	171

Note: "Average speed" indicates the average number of seconds it took respondents to answer four political knowledge questions located early in the questionnaire. "Non-attitudes rate" is the percentage of questions where respondents reported "I don't know" or did not provide a response, among four political knowledge questions, three questions on attitudes toward public deliberation, and a check-all-that-apply question about participation in political activities. "Straightlining" gives the proportion of respondents choosing the same option ("Support", "Oppose", "I'm indifferent", or "I don't know") on questions presented on a grid about support for six national policies. "Preference consistency" gives the proportion of respondents reporting incomplete and intransitive sets of strict pairwise preferences over a set of policy options aimed at preventing legislative gridlock in the state legislature. Since we focus on strict orderings, respondents selecting "I'm indifferent" for any pair of policies are coded as having incomplete preferences. The proportion of respondents reporting intransitive preferences is calculated among those who report complete strict preferences only. Bold numbers indicate a statistically significant difference in means (at a 95% confidence level) between respondents that pass and fail.

Table B.4: Linear regression analysis of overall political knowledge

	Model 1		**Model 2**		**Model 3**	
Coefficient	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.87	0.03	0.76	0.12	0.48	0.12
Fail TQ 1	-0.31	0.06	-0.22	0.06		
Fail TQ 2	-0.74	0.09	-0.53	0.09		
Fail TQ 3	-0.7	0.1	-0.49	0.1		
Education			0.3	0.02	0.31	0.02
Age			0.01	0	0.01	0
Female			-0.38	0.05	-0.38	0.05
SoCal (excl. LA)			-0.21	0.07	-0.2	0.07
SoCal (LA)			-0.11	0.07	-0.13	0.07
Central/Southern Farm			-0.3	0.09	-0.31	0.09
North and Mountain			0.06	0.11	0.04	0.11
Central Valley			0.07	0.1	0.06	0.1
Adjusted R2		0.04		0.14		0.12
N		2695		2695		2695

Note: The table presents linear regression results. Dependent variable: 0-4 political knowledge scale. Geographical area used as baseline for the region indicator: Bay Area. The F-statistic for Model 2 relative to Model 1 is 41.8, and the one for Model 3 relative to Model 2 is 19.4. 30 respondents are dropped from all linear regressions due to missing region variable.

Table B.5: Linear regression analysis of political participation

	Model 1		Model 2		Model 1	
Coefficient	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.88	0.05	0.12	0.21	-0.21	0.2
Fail TQ 1	-0.51	0.1	-0.34	0.1		
Fail TQ 2	-0.69	0.15	-0.27	0.15		
Fail TQ 3	-1.27	0.17	-0.88	0.17		
Education			0.54	0.04	0.56	0.04
Age			0.02	0	0.02	0
Female			-0.04	0.08	-0.04	0.08
SoCal (excl. LA)			0.06	0.12	0.06	0.12
SoCal (LA)			0.01	0.12	0	0.12
Central/Southern Farm			0.21	0.15	0.21	0.15
North and Mountain			0.25	0.19	0.23	0.19
Central Valley			0.13	0.17	0.14	0.17
Adjusted R2		0.03		0.11		0.1
N		2695		2695		2695

Note: The table presents linear regression results. Dependent variable: 0-12 political participation scale. Geographical area used as baseline for the region indicator: Bay Area. The F-statistic for Model 2 relative to Model 1 is 34.0, and the one for Model 3 relative to Model 2 is 11.7. 30 respondents are dropped from all linear regressions due to missing region variable.

Table B.6: Linear regression analysis of strength of ideological leanings

	Model 1		**Model 2**		**Model 3**	
Coefficient	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.99	0.05	2.37	0.22	1.94	0.21
Fail TQ 2	-0.83	0.14	-0.68	0.14		
Fail TQ 3	-1.24	0.16	-1.09	0.16		
Education			0.19	0.04	0.23	0.04
Age			0.01	0	0.01	0
Female			0.09	0.08	0.1	0.08
SoCal (exc LA)			-0.53	0.13	-0.54	0.13
SoCal (LA)			-0.29	0.13	-0.33	0.13
Central/Southern Farm			-0.7	0.16	-0.73	0.16
North and Mountain			-0.65	0.21	-0.69	0.21
Central Valley			-0.17	0.18	-0.2	0.18
Adjusted R2		0.04		0.07		0.04
N		2097		2097		2097

Note: The table presents linear regression results. Dependent variable: 0-6 ideology strength scale (absolute value of ideology scale presented in Figure 3). Geographical area used as baseline for the region indicator: Bay Area. The F-statistic for Model 2 relative to Model 1 is 8.648, and the one for Model 3 relative to Model 2 is 31.193. Total number of observations: 2,097. 598 respondents are dropped due to missing responses to at least one of the six policy questions used to construct the ideology scale (in addition to 30 missing region variable).

Table B.7: Organizations included in double-list experiment

Nonsensitive, List A:
Californians for Disability Rights (organization advocating for people with disabilities)
California National Organization for Women (organization advocating for women's equality and empowerment)
American Family Association (organization advocating for pro-family values)
American Red Cross (humanitarian organization)

Nonsensitive, List B:
American Legion (veterans service organization)
Equality California (gay and lesbian advocacy organization)
Tea Party Patriots (conservative group supporting lower taxes and limited government)
Salvation Army (charitable organization)

Sensitive, X condition
Organization X (organization advocating for immigration reduction and measures against undocumented immigration)

Sensitive, Y condition
Organization Y (citizen border patrol group combating undocumented immigration)

Note: The table lists items displayed to respondents in the double list experiment on support for anti-immigrant organizations. Displayed information included both organization names and descriptions. The table gives organizations listed in list A; including non-sensitive items (displayed to all respondents) and sensitive items X and Y (displayed to respondents in the corresponding treatment group for list A). The table also gives organizations listed in list B; including non-sensitive items (displayed to all respondents) and sensitive items X and Y (displayed to respondents in the corresponding treatment group for list B).

Table B.8: Number of respondents in each experimental condition

	Experimental condition		
Sensitive item	Control A – Treatment B	Treatment A – Control B	Total
Organization X	525 (24.4%)	545 (25.4%)	1,070 (49.8%)
Organization Y	542 (25.2%)	537 (25.0%)	1,079 (50.2%)
Total	1,067 (49.6%)	1,082 (50.4%)	2,149 (100%)

Note: The table shows the number and percentage of respondents assigned to each combination of experimental condition (columns) and sensitive item displayed to respondents in the treatment group (rows).

Table B.9: Attentiveness and difference-in-means estimates

		Mean response (control)	Mean response (Org. X)	Mean response (Org. Y)	Diff.-in-means (Org. X)	Diff.-in-means (Org. Y)
List A	Attentive	2.32	2.68	2.54	0.36	0.22
	Std. error	(0.04)	(0.07)	(0.07)	(0.09)	(0.09)
	Inattentive	1.93	1.9	2.03	-0.03	0.1
	Std. error	(0.11)	(0.17)	(0.17)	(0.20)	(0.20)
	Difference	0.39	0.78	0.51	0.39	0.13
	Std. error	(0.12)	(0.18)	(0.18)	(0.21)	(0.22)
List B	Attentive	2.17	2.42	2.45	0.25	0.28
	Std. error	(0.04)	(0.07)	(0.06)	(0.08)	(0.08)
	Inattentive	1.58	2.18	2.11	0.6	0.53
	Std. error	(0.09)	(0.17)	(0.18)	(0.19)	(0.20)
	Difference	0.58	0.24	0.34	-0.35	-0.25
	Std. error	(0.10)	(0.18)	(0.19)	(0.21)	(0.22)
List B – List A:						
	Attentive				-0.11	0.06
	Std. error				(0.13)	(0.14)
	Inattentive				0.62	0.43
	Std. error				(0.34)	(0.35)

Note: The first three columns show mean responses under control, X-treatment, and Y-treatment conditions for attentive and inattentive respondents for list A (top) and list B (middle). The last two columns show the difference in mean responses under treatment (X or Y) and under control for attentives and inattentives in list A and in list B. The differences between attentives and inattentives in terms of mean responses and difference-in-means estimates are also calculated. The bottom section ("List B – List A") shows the differences across lists in terms of difference-in-means estimates for attentives and inattentives respectively. Bootstrapped standard errors are provided between parentheses.

Table B.10: Number of selected items in double-list experiment

List A	Attentive			Inattentive		
Response	Control	Org. X	Org. Y	Control	Org. X	Org. Y
0	0.13	0.12	0.12	0.25	0.31	0.27
1	0.16	0.11	0.15	0.14	0.13	0.14
2	0.22	0.21	0.2	0.27	0.25	0.21
3	0.25	0.24	0.27	0.11	0.11	0.16
4	0.24	0.15	0.12	0.23	0.09	0.1
5		0.17	0.14		0.12	0.12
Observations	880	444	426	187	101	111
Mean	2.32	2.68	2.54	1.93	1.9	2.03
Std. deviation	1.34	1.57	1.54	1.47	1.71	1.69

List B	Attentive			Inattentive		
Response	Control	Org. X	Org. Y	Control	Org. X	Org. Y
0	0.1	0.13	0.1	0.28	0.22	0.2
1	0.19	0.15	0.16	0.2	0.15	0.18
2	0.26	0.22	0.24	0.29	0.22	0.3
3	0.34	0.26	0.27	0.12	0.19	0.08
4	0.11	0.15	0.16	0.11	0.08	0.11
5		0.09	0.08		0.14	0.13
Observations	870	425	455	212	100	87
Mean	2.17	2.42	2.45	1.58	2.18	2.11
Std. deviation	1.16	1.47	1.39	1.31	1.67	1.63

Note: Respondents fall into six categories according to attentiveness and the condition they are assigned to (control, X-treatment or Y-treatment). The table shows for each category the distribution of number of items selected by respondents, the standard deviation as well as the average. The table also shows the difference-in-means estimates for organization X and Y for attentives and inattentives respectively, and the difference in these estimates between attentives and inattentives. Results for list A and B are displayed on the top and on the bottom respectively.

Table B.11: Transition matrices between two lists

X-treatment for list A (row) and control for list B (column):
Attentives:

	0	1	2	3	4	N
0	0.6	0.19	0.15	0.06	0	53
1	0.16	0.45	0.25	0.12	0.02	51
2	0.03	0.27	0.38	0.24	0.07	94
3	0.02	0.15	0.25	0.53	0.05	107
4	0	0.05	0.18	0.48	0.29	65
5	0.01	0.03	0.18	0.43	0.35	74

Inattentives:

	0	1	2	3	4	N
0	0.84	0.1	0.06	0	0	31
1	0.08	0.62	0.23	0	0.08	13
2	0.04	0.4	0.32	0.12	0.12	25
3	0.09	0.09	0.73	0.09	0	11
4	0	0	0.22	0.78	0	9
5	0	0.08	0.25	0.08	0.58	12

Y-treatment for list A (row) and control for list B (column):
Attentives:

	0	1	2	3	4	N
0	0.53	0.37	0.08	0	0.02	51
1	0.12	0.52	0.22	0.12	0.02	65
2	0.05	0.26	0.43	0.24	0.02	84
3	0	0.05	0.39	0.51	0.04	114
4	0	0.08	0.19	0.58	0.15	53
5	0.05	0.03	0.14	0.41	0.37	59

Inattentives:

	0	1	2	3	4	N
0	0.77	0.03	0.1	0.07	0.03	30
1	0.19	0.56	0.19	0	0.06	16
2	0.13	0.22	0.52	0.13	0	23
3	0.06	0.17	0.56	0.11	0.11	18
4	0	0.09	0.36	0.18	0.36	11
5	0	0.08	0.23	0.31	0.38	13

Transition matrices between two lists (continued)

Control for list A (row) and X-treatment for list B (column):
Attentives:

	0	1	2	3	4	5	N
0	0.64	0.16	0.1	0.07	0.02	0.02	58
1	0.11	0.41	0.32	0.07	0.07	0.01	71
2	0.07	0.14	0.37	0.3	0.1	0.03	103
3	0.01	0.06	0.12	0.46	0.28	0.06	93
4	0.02	0.06	0.16	0.27	0.23	0.26	100

Inattentives:

	0	1	2	3	4	5	N
0	0.68	0.14	0.07	0.04	0	0.07	28
1	0.07	0.43	0.36	0.14	0	0	14
2	0.04	0.2	0.44	0.32	0	0	25
3	0.11	0	0.33	0.22	0.33	0	9
4	0	0	0.04	0.25	0.21	0.5	24

Control for list A (row) and Y-treatment for list B (column):
Attentives:

	0	1	2	3	4	5	N
0	0.56	0.29	0.07	0.04	0.04	0	55
1	0.11	0.39	0.29	0.1	0.06	0.06	70
2	0.05	0.2	0.37	0.25	0.11	0.02	92
3	0.01	0.08	0.25	0.42	0.18	0.06	123
4	0	0.02	0.18	0.32	0.29	0.19	115

Inattentives:

	0	1	2	3	4	5	N
0	0.83	0.11	0.06	0	0	0	18
1	0.08	0.69	0.08	0.08	0	0.08	13
2	0	0.12	0.65	0.15	0.04	0.04	26
3	0.09	0	0.36	0.09	0.45	0	11
4	0	0.11	0.16	0.05	0.21	0.47	19

Note: The matrices are the transition matrices for respondents' choices between two lists, separately for attentive and inattentive respondents in each experimental condition.

Table B.12: Attentiveness and difference-in-means estimates for Eady (2017)

	Mean response (control)	Mean response (treated)	Diff.-in-means
Pass	1.61	2.49	0.88
Std. error	(0.01)	(0.01)	(0.01)
Fail	1.51	2.26	0.75
Std. error	(0.03)	(0.04)	(0.05)
Difference	0.1	0.23	0.13
Std. error	(0.04)	(0.04)	(0.05)

Note: The table shows mean responses under control and under treatment for respondents who passed and failed the screener question in Eady (2017). The last column shows the difference in mean responses under control and under treatment for respondents who passed and failed the screener question. The differences between for respondents who passed and failed the screener question in terms of mean responses and difference-in-means estimates are also calculated and reported in the bottom row.

Table B.13: Linear regression analysis of support for anti-immigrant organizations

	Term	Model 1 Estimate	SE	Model 2 Estimate	SE	Model 3 Estimate	SE
Condition	Control						
	Organization X	0.38	0.08	0.73	0.38	0.71	0.38
	Organization Y	0.22	0.08	-0.19	0.37	-0.17	0.36
Trap	Pass						
	Fail	-0.35	0.12	-0.38	0.12	-0.28	0.12
List	List A						
	List B	-0.15	0.07	-0.14	0.07	-0.14	0.07
Interactions	Org. X x Fail	-0.42	0.2	-0.41	0.2	-0.39	0.2
	Org. Y x Fail	-0.13	0.19	-0.12	0.2	-0.11	0.19
	Org. X x List B	-0.13	0.12	-0.16	0.12	-0.15	0.11
	Org. Y x List B	0.06	0.12	0.05	0.12	0.04	0.11
	Fail x List B	-0.23	0.16	-0.24	0.16	-0.24	0.16
	Org. X x Fail x List B	0.84	0.28	0.88	0.28	0.83	0.27
	Org. Y x Fail x List B	0.39	0.28	0.43	0.28	0.44	0.27
	Intercept	2.31	0.05	2.64	0.21	2.38	0.21
Demographics		No		Yes		Yes	
Additional Controls		No		No		Yes	
Adj. R-squared			0.03		0.04		0.08
N		2096 x 2		2096 x 2		2096 x 2	

Note: The table shows linear regression results for three specifications. The dependent variable is the number of items selected by a respondent for all linear regression models. Independent variables for the baseline specification are condition dummies (control, X-treatment or Y-treatment), attentiveness dummy (pass or fail the trap questions), list dummy (list A or list B) and all interaction terms. Model 2 also includes demographic variables (gender, education, age and region) and their interactions with treatment dummies. Model 3 further includes three additive measures of political knowledge, political participation, and ideological leaning. The baseline demographic group for the last two linear regressions is female, without a high school degree, aged below 25, and from Bay Area. Standard errors are clustered at respondent level. 53 respondents are dropped due to missing values for Model 3 variables.

Appendix C

APPENDIX TO CHAPTER IV

Details about Administrative Data

Administrative Data

Given the importance of administrative data for the work we report in this paper, here we provide additional details for interested readers. Administrative data, like voter registration records, have long been used by political science researchers. Some use voter registration records to provide contact information for field experiments (e.g., Gerber and Green 2000), as sampling frames to improve the accuracy of electoral polling (e.g. Green and Gerber 2006), or to study the misreporting of voter turnout in surveys (e.g., Ansolabehere and Hersch 2017). These studies all assume that the administrative records they use are accurate.

In the United States, states are now generally required to have a statewide voter registration database due to the *Help America Vote Act* (HAVA, 2002). These statewide voter registration datasets are used for many purposes, other than academic research. Campaigns and political parties use them for voter mobilization and persuasion (Hersch 2015). State, county, and municipal election officials use these data for voter information activities, to send vote-by-mail ballots, to allocate resources for in-person voting, and to authenticate eligible voters during an election. Clearly election officials have important incentives to keep these data as accurate as possible.

The procedures that states can use to maintain the accuracy of their voter registration datasets is regulated by the *National Voter Registration Act* (NVRA, 1993). Each state then issues it's own mechanisms for compliance with NVRA, and California's rules are provided by the Secretary of State (`https://www.sos.ca.gov/elections/voter-registration/nvra/laws-standards/nvra-manual`). In general, the NVRA regulations provide guidance about how potentially ineligible voters can be moved to an inactive status or removed from the file, usually based on changes like moving outside of the jurisdiction or death. In our paper we use data from Orange County, and their registration list maintenance procedures are provided online (`https://www.ocvote.com/registration/maintaining-an-accurate-voter-list`).

As has been noted in recent research using these administrative data from Orange County (Kim, Schneider, and R Michael Alvarez, 2020), in a jurisdiction of this size, there will be record changes, new records, and removed records. However, this recent research has shown that these changes in the administrative data are relatively small in relation to the overall file, and there is no research that we are aware of that would indicate that file maintenance or inaccuracies in the administrative data more generally should have any effect on our estimate of the quantities we examine in our paper. In particular, the components of the voter registration data that are most subject to change and possible error are the fields with voter identifying information — their name, address, and other voter-provided information. The information from the administrative files we used in our paper, each registered voter's participation in current and past elections, comes from the county's election management system and we have every reason to believe that these components of the administrative record are accurate.

By focusing our research on a large and diverse county, we argue that our results should generalize. Orange County, California, is a very large and diverse election jurisdiction located in Southern California. With about 3 million residents, in the 2018 general election Orange County had just over 2 million citizens eligible to vote, with nearly 1.6 million registered voters.[1]. It is one of the largest election jurisdictions in the United States, closely divided in partisan registration: in 2018, the county had 541,665 registered Republicans, 523,624 registered Democrats, and 429,675 registered No Party Preference voters.

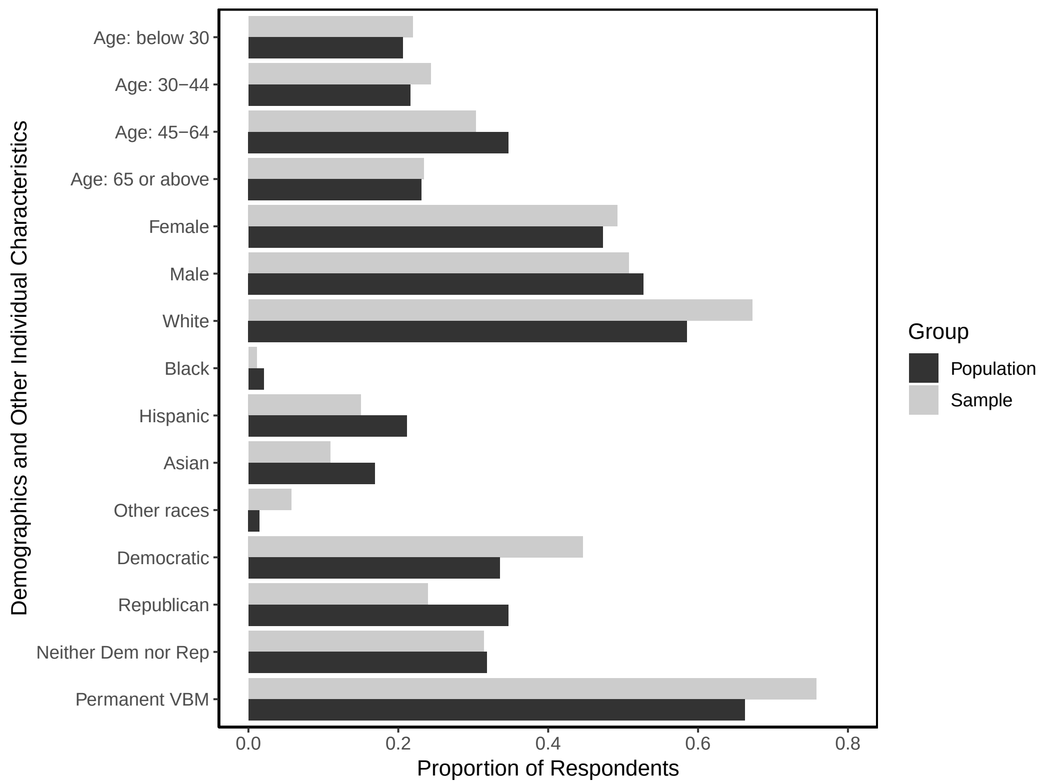
Details about the Survey

Among 6,952 respondents who completed our survey, we are able to determine basic demographic information for 6,912 respondents. Figure C.1 shows the demographic compositions and the distributions of other individual characteristics for our survey respondents and the population of Orange County registered voters before the November 2018 general election.

Registered voters of different ages and genders are well represented in our survey sample. There are slightly more (2.7%) respondents between 30 and 44 years old, fewer (4.4%) respondents from age group 45-64, and fewer (2.0%) female respondents, compared to the population of registered voters. Our survey sample

[1] https://elections.cdn.sos.ca.gov/sov/2018-general/sov/02-county-voter-reg-stats-by-county.pdf

Figure C.1: Respondent Composition of the Survey



exhibits more imbalances in race/ethnicity and party registration. More white voters and fewer Hispanic or Latino and Asian American voters participated in the survey (by 8.7%, 6.1%, and 6.0% respectively). Meanwhile, while 34.7% of voters registered with the Republican party in Orange County, only 24.0% of those who completed our survey are Republican voters. On the other hand, 44.6% of respondents are Democratic voters, compared to 33.5% in the population. The disparities in terms of race/ethnicity and party registration in our survey sample are expected given our knowledge about survey participation in general and consistent with other surveys with voluntary participation. Finally, 66.2% of registered voters in Orange County are permanent vote-by-mail voters, whereas the percentage is 75.8% in our sample; the distribution of cities of residence for our sample tracks the population well.

Of course, we could use calibration weighting to adjust our sample to population parameters (Caughey et al. 2020). We decided here to use unweighted data, as we did not want to make population estimates specific to Orange County. That said, our unweighted estimates should generalize to other similar large-scale surveys and polls. Weighting to Orange County population parameters would not change the general results we see in our analysis, and as Orange County is a large and diverse population, our results should apply to surveys and polls of other large and diverse populations.

Attention Check Questions

Main IRI: Do you support or oppose any of the following proposals for new ways of voting or conducting elections?

[For each of the following statements, respondents can select *Support strongly*, *Support somewhat*, *Neither support nor oppose*, *Oppose somewhat*, *Oppose strongly*]

- Allow absentee voting over the Internet
- Run all elections by mail
- Allow people to register on Election Day at the polls
- Require all people to show government issued photo ID when they vote
- Require electronic voting machines to print a paper backup of the ballot
- *For half of the respondents*: For survey quality control, select Oppose strongly
- *For half of the respondents*: For survey quality control, select Support strongly

Note: To minimize order effects, response options are presented in ascending order for half of the respondents, and in descending order for the other half of the respondents.

Main IMC: When a big news story breaks people often go online to get up-to-the-minute details on what is going on. People differ in which websites they trust to get this information. For survey quality control, please ignore the question and select ABC News and Reuters as your two answers.

When there is a big news story, which websites you would visit first?

- New York Times website
- The Drudge Report
- The Associated Press (AP) website
- Huffington Post
- Google News
- NBC News website
- Washington Post website
- ABC News website
- National Public Radio (NPR) website
- CNN.com
- CBS News website
- USA Today website
- FoxNews.com
- Reuters website
- New York Post Online
- MSNBC.com
- Yahoo! News
- None of these websites

Note: The position of prompt *For survey quality control, please ignore the question and select ABC News and Reuters as your two answers* is randomized between the middle of the question and the end of the question.

Questions on Turnout and Mode of Voting

Note: Turnout Question 1 (2018 General Turnout), Turnout Question 2a, and Mode of Voting Question 1 (2018 General Mode of Voting) appear right before the first attention check, while Turnout Question 2b, Turnout Question 3 (Turnout in Previous Elections) and Mode of Voting Question 2 (Mode of Voting in Previous Elections) appear right after the second attention check.

Turnout Question 1: Which of the following statements best describes you?

- I did not vote in the election this November
- I thought about voting this time, but didn't
- I usually vote, but didn't this time
- I tried to vote, but was not allowed to when I tried
- I tried to vote, but it ended up being too much trouble
- I definitely voted in the November General Election

[If the respondent chose one of the last three response options in Turnout Question 1, they receive the following two questions.]

Turnout Question 2a Was this your first time voting, or have you voted in elections before? [Randomize first two options.]

- I am a first time voter
- I have voted in elections before
- I don't know

Mode of Voting Question 1 How did you vote this election?

- Voted in person on Election Day (at a polling place or precinct)
- Voted in person before Election Day
- Voted by mail or absentee ballot by mail
- I don't know

[If the respondent chose one of the first three response options in Turnout Question 1, they receive the following question.]

Turnout Question 2b Have you voted in elections before? [Randomize first two options.]

- Yes
- No
- I don't know

[If the respondent chose I have voted in elections before in Turnout Question 2a or Yes in Turnout Question 2b, they receive the following questions.]

Turnout Question 3 Earlier you mentioned that you have voted in elections before. Did you vote in the following general elections? [For each of the following statements, respondents can select *Yes*, *No*, or *I don't remember*.]

- November 2016 Presidential General Election
- November 2014 Midterm General Election
- November 2012 Presidential General Election

. . . in the following primary elections?

- June 2018 Statewide Primary Election
- June 2016 Presidential Primary Election

Mode of Voting Question 2 How did you vote in these general elections you voted in? [For each of the following statements, respondents can select *Voted in person*, *Voted by mail or absentee ballot by mail*, or *I don't know*. Only those elections that a respondent answers Yes to in Turnout Question 3 are included.]

- November 2016 Presidential General Election
- November 2014 Midterm General Election
- November 2012 Presidential General Election

How did you vote in the primary elections you voted in?

- June 2018 Statewide Primary Election
- June 2016 Presidential Primary Election

Table C.1: Demographics and Passage of Attention Checks (Logistic Regression)

	IRI		IMC	
	Avg. ME	Std. Error	Avg. ME	Std. Error
Below 30				
30 - 44	-0.03	0.01	-0.03	0.02
45 - 64	-0.07	0.01	-0.11	0.02
65 or above	-0.12	0.01	-0.16	0.02
Female				
Male	0.00	0.01	-0.02	0.01
Rather not say	-0.04	0.05	-0.08	0.06
Self-describe	0.02	0.05	-0.06	0.07
HS or less				
Some college	0.06	0.03	0.13	0.03
2-year college	0.03	0.03	0.16	0.04
4-year college	0.08	0.03	0.19	0.03
Postgraduate	0.11	0.03	0.24	0.03
White				
Hispanic	-0.12	0.02	-0.15	0.02
Asian	-0.11	0.02	-0.09	0.02
Black	-0.19	0.05	-0.17	0.06
Other races	-0.08	0.02	-0.23	0.03

Table C.2: Accuracy of Self-Reported Birth Year, City of Residence, and Voter Registration

	Incorrect	Correct	Skip	N
Birth Year:				
Fail IRI	6.3%	92.9%	0.8%	999
Pass IRI	3.4%	96.0%	0.6%	5648
Fail IMC	4.8%	94.8%	0.4%	3072
Pass IMC	3.2%	96.2%	0.6%	3618
City of Residence:				
Fail IRI	8.4%	90.9%	0.6%	971
Pass IRI	6.8%	92.8%	0.4%	5474
Fail IMC	7.9%	91.7%	0.4%	2994
Pass IMC	6.8%	92.9%	0.3%	3496
Voter Registration:				
Fail IRI	8.5%	88.0%	3.4%	986
Pass IRI	6.9%	90.9%	2.2%	5608
Fail IMC	8.3%	88.7%	3.0%	3037
Pass IMC	6.3%	92.0%	1.8%	3595

Table C.3: Inattentive Respondents Are More Likely to Misreport Turnout

	All	Fail IRI	Pass IRI	Fail IMC	Pass IMC
2018 General	0.033	0.041	0.031	0.039	0.028
	(0.002)	(0.006)	(0.002)	(0.003)	(0.003)
2018 Primary	0.137	0.184	0.129	0.174	0.106
	(0.004)	(0.013)	(0.005)	(0.007)	(0.006)
2016 General	0.073	0.102	0.067	0.09	0.06
	(0.004)	(0.011)	(0.004)	(0.006)	(0.004)
2016 Primary	0.212	0.272	0.203	0.239	0.186
	(0.006)	(0.017)	(0.007)	(0.009)	(0.008)
2014 General	0.236	0.275	0.23	0.265	0.213
	(0.007)	(0.019)	(0.008)	(0.011)	(0.009)
2012 General	0.135	0.17	0.13	0.144	0.126
	(0.006)	(0.016)	(0.006)	(0.009)	(0.008)

Table C.4: Inattentive Respondents Are More Likely to Misreport Mode of Voting

	All	Fail IRI	Pass IRI	Fail IMC	Pass IMC
2018 General	0.055	0.072	0.053	0.061	0.053
	(0.003)	(0.008)	(0.003)	(0.004)	(0.004)
2018 Primary	0.091	0.105	0.089	0.095	0.085
	(0.005)	(0.014)	(0.005)	(0.007)	(0.006)
2016 General	0.119	0.135	0.118	0.129	0.111
	(0.005)	(0.014)	(0.005)	(0.008)	(0.007)
2016 Primary	0.116	0.133	0.115	0.118	0.114
	(0.006)	(0.017)	(0.007)	(0.009)	(0.008)
2014 General	0.124	0.123	0.126	0.126	0.124
	(0.007)	(0.018)	(0.008)	(0.011)	(0.009)
2012 General	0.161	0.199	0.157	0.165	0.155
	(0.007)	(0.021)	(0.008)	(0.011)	(0.01)

Table C.5: Respondent Attention Is Positively Correlated with Validated Turnout

	All	Fail IRI	Pass IRI	Fail IMC	Pass IMC
2018 General	0.952	0.939	0.955	0.944	0.958
	(0.003)	(0.008)	(0.003)	(0.004)	(0.003)
2018 Primary	0.707	0.638	0.718	0.683	0.727
	(0.006)	(0.016)	(0.006)	(0.009)	(0.008)
2016 General	0.894	0.865	0.9	0.878	0.907
	(0.004)	(0.012)	(0.005)	(0.007)	(0.005)
2016 Primary	0.688	0.639	0.698	0.665	0.71
	(0.007)	(0.018)	(0.007)	(0.01)	(0.009)
2014 General	0.622	0.561	0.635	0.601	0.64
	(0.007)	(0.02)	(0.008)	(0.011)	(0.01)
2012 General	0.835	0.789	0.84	0.83	0.84
	(0.006)	(0.017)	(0.007)	(0.009)	(0.008)

Table C.6: Respondent Attention Is *Not* Correlated with Validated Mode of Voting

	All	Fail IRI	Pass IRI	Fail IMC	Pass IMC
2018 General	0.681	0.666	0.685	0.678	0.689
	(0.006)	(0.015)	(0.006)	(0.009)	(0.008)
2018 Primary	0.661	0.638	0.664	0.666	0.657
	(0.007)	(0.021)	(0.008)	(0.011)	(0.01)
2016 General	0.62	0.625	0.622	0.618	0.626
	(0.007)	(0.019)	(0.008)	(0.011)	(0.01)
2016 Primary	0.613	0.619	0.611	0.611	0.61
	(0.009)	(0.023)	(0.009)	(0.013)	(0.012)
2014 General	0.603	0.593	0.601	0.613	0.598
	(0.01)	(0.027)	(0.011)	(0.015)	(0.013)
2012 General	0.556	0.559	0.555	0.547	0.558
	(0.009)	(0.024)	(0.01)	(0.014)	(0.012)

Table C.7: Dropping Inattentive Respondents Does *Not* Reduce Bias in Turnout Estimates

	All	Pass IRI	Pass IMC
Bias:			
2018 General	0.021	0.025	0.03
2018 Primary	0.094	0.098	0.092
2016 General	0.062	0.064	0.068
2016 Primary	0.196	0.194	0.199
2014 General	0.236	0.238	0.239
2012 General	0.115	0.118	0.12
Std. Error:			
2018 General	0.002	0.002	0.002
2018 Primary	0.005	0.006	0.007
2016 General	0.003	0.003	0.004
2016 Primary	0.005	0.005	0.006
2014 General	0.006	0.006	0.008
2012 General	0.004	0.004	0.005
RMSE:			
2018 General	0.021	0.025	0.03
2018 Primary	0.094	0.099	0.092
2016 General	0.062	0.064	0.068
2016 Primary	0.196	0.194	0.199
2014 General	0.236	0.238	0.24
2012 General	0.115	0.118	0.121

Table C.8: Dropping Inattentive Respondents Reduces Bias in Voting-by-Mail Estimates

	All	Pass IRI	Pass IMC
Bias:			
2018 General	-0.055	-0.048	-0.044
2018 Primary	-0.049	-0.041	-0.043
2016 General	-0.07	-0.062	-0.061
2016 Primary	-0.051	-0.048	-0.045
2014 General	-0.056	-0.052	-0.047
2012 General	-0.093	-0.094	-0.094
Std. Error:			
2018 General	0.006	0.006	0.008
2018 Primary	0.007	0.008	0.01
2016 General	0.007	0.008	0.01
2016 Primary	0.008	0.009	0.011
2014 General	0.009	0.01	0.012
2012 General	0.009	0.01	0.012
RMSE:			
2018 General	0.055	0.049	0.044
2018 Primary	0.05	0.041	0.044
2016 General	0.07	0.062	0.061
2016 Primary	0.052	0.049	0.047
2014 General	0.056	0.053	0.048
2012 General	0.093	0.094	0.095