An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions[*]

Bryn Rosenfeld[†] Kosuke Imai[‡] Jacob Shapiro[§]

June 11, 2014

Abstract

When studying sensitive issues such as corruption, prejudice, and sexual behavior, researchers have increasingly relied upon indirect questioning techniques to mitigate such known problems of direct survey questions as under-reporting and nonresponse. However, there have been surprisingly few empirical validation studies of these indirect techniques, because the information required to verify the resulting estimates is often difficult to access. This paper reports findings from the first comprehensive validation study of indirect methods. We estimate whether people voted for an anti-abortion referendum held during the 2011 Mississippi General Election using direct questioning and three popular indirect methods: list experiment, endorsement experiment, and randomized response. We then validate these estimates against the official election outcome. While direct questioning leads to significant under-estimation of sensitive votes against the referendum, these survey techniques yield estimates much closer to the actual vote count, with endorsement experiment and randomized response yielding least bias.

Key Words: endorsement experiments, indirect questioning, item count technique, list experiment, randomized response, social desirability bias

1 Introduction

Many of the topics social scientists study are sensitive and private in nature. When studying such issues as corruption, prejudice, and sexual behavior, obtaining accurate measures of citizens' sensitive attitudes and behavior poses a serious methodological challenge. Direct survey questions of these topics often lead to a substantial amount of under-reporting and non-response. To reduce possible biases due to social desirability and missing data, researchers increasingly rely upon several indirect questioning techniques such as list experiment (also known as item count technique or unmatched count technique), endorsement experiment, and randomized response (e.g., Blair *et al.*, 2013; Gonzalez-Ocantos *et al.*, 2012; Krumpal, 2012; Gingerich, 2010; Kuklinski *et al.*, 1997; Lyall *et al.*, 2013). As their applications increase, new methodologies have been developed to analyze responses to these indirect questioning techniques and statistically estimate truthful responses to sensitive questions (e.g., Corstange, 2009; Gingerich, 2010; Bullock *et al.*, 2011; Imai, 2011; Blair and Imai, 2012; Glynn, 2013).

Despite the increasing popularity of these survey methodologies, the difficulty of gaining access to suitable sensitive information means that there are few empirical validation studies. Certainly at the individual-level and even at low levels of aggregation, sensitive records are usually confidential. While in a handful of exceptional cases researchers have validated the responses of direct questions against official records (e.g., Folsom, 1974; Junger, 1989; van der Heijden *et al.*, 2000; Hessing *et al.*, 1988; Elffers *et al.*, 1992), validation studies of indirect questioning techniques remain relatively rare. For example, a review paper by Lensvelt-Mulders *et al.* (2005) lists only six validation studies of the randomized response technique (Horvitz *et al.*, 1967; Kulka *et al.*, 1981; Lamb and Stem, 1978; Locander *et al.*, 1976; Tracy and Fox, 1981; van der Heijden *et al.*, 2000),[1] despite the fact that the method has been in use for almost half a century since the pioneering work of Warner (1965). Furthermore, to the best of our knowledge, there have been no validation studies of either list experiments or endorsement experiments.

---

[1] In addition, we also found another validation study by Wolter and Preisendorfer (2013) that was published more recently.

In this paper, we report findings from the first comprehensive validation study to directly assess the empirical performance of four commonly used survey methods for measuring sensitive attitudes and behavior: direct questioning, list experiment, endorsement experiment, and randomized response. As in other validation studies, we compare the estimates of a sensitive trait based on various survey methodologies to the corresponding truth. To do so, we exploit the official election outcome for a sensitive anti-abortion referendum held during the 2011 Mississippi General Election. Although official records do not reveal the vote choice of individuals in our sample, the Mississippi Secretary of State's county recapitulation reports provide the true vote share at low levels of aggregation, i.e., counties. We sample only those who actually turned out in this election using the public voter history records and ask them how they voted on the referendum. This allows us to directly evaluate how well each survey methodology recovers ground truth.

We find that direct questioning leads to significant under-estimation of casting a "no" vote on the referendum, which is the socially undesirable behavior in this context, by more than 20 percentage points in most counties. In contrast, all three indirect techniques provide estimates much closer to the actual vote count. The endorsement experiment and the randomized response yield least bias, but the estimates based on the endorsement experiment (with a single item) are noisier than other indirect questioning methods. Across 19 counties, we find that the bias for these two methods can be as little as 10% of the bias for the direct question. The list experiment has a smaller bias than direct questioning and is less noisy than the endorsement experiment, but the magnitude of its bias exceeds that of the other two indirect questioning methods. Thus, the randomized response appears to outperform both the list and endorsement experiments. This result contradicts recent studies, which reached skeptical conclusions about randomized response methods without access to the information necessary for validation (Coutts and Jann, 2011; Holbrook and Krosnick, 2010a), but is more consistent with the result of a comprehensive meta-analysis conducted by Lensvelt-Mulders *et al.* (2005).

These findings rest on a more robust scientific ground than numerous existing studies, which simply compare estimates from multiple measurement strategies in the absence of the true sensitive

information (e.g., Anderson *et al.*, 2007; Dalton *et al.*, 1994; Holbrook and Krosnick, 2010a,b; LaBrie and Earleywine, 2000; Tsuchiya *et al.*, 2007). Under what Tourangeau and Yan (2007) call the "more is better" assumption, these comparative studies consider the method which produces the largest (smallest) estimate of the sensitive socially undesirable (desirable) behavior to be the most accurate one. Clearly, this assumption may not be warranted in some cases and even where it is, such purely comparative designs are unable to quantify the absolute magnitude of bias. Validation studies like ours overcome these problems by comparing estimates directly against true information.

In the reminder of the paper, we first discuss our experiment and review various approaches to measuring sensitive attitudes and behaviors (Section 2). We then describe the statistical approaches we employ to analyze the responses to indirect questioning (Section 3). In Section 4, we report our empirical findings, summarizing the relative performance of the various measure in terms of both non-response rates and bias. Finally, we conclude by discussing the implications of our results for applied work and outlining potential avenues for future research (Section 5).

2 The Design of the Mississippi Validation Study

In this section, we discuss the motivation for basing our study on a sensitive anti-abortion referendum, know as the "personhood amendment", on the 2011 Mississippi General Election ballot. We also provide a brief review of four common survey methods for eliciting sensitive political attitudes and behaviors: direct questioning, list experiments, endorsement experiments, and randomized response methods. Each of the three indirect questioning techniques prevents the researcher from identifying any individual respondent's truthful position in a distinct way.

- The list experiment masks individual responses through *aggregation*. It asks respondents questions about a set of actions or views at once rather than the sensitive one in isolation. To assess the prevalence of sensitive attitudes and behavior, the researcher randomizes whether the sensitive item of interest is added to the list of control items.

- The endorsement experiment obscures individual responses by exploiting *evaluation bias* in human judgement. This draws on a rich literature in psychology, which demonstrates that

4

people tend to evaluate identical objects positively (negatively) when paired with favorable (unfavorable) entities. It asks respondents non-sensitive questions but randomizes whether these questions are paired with the sensitive object.

- The randomized response method obscures individual responses by adding *random noise*. It asks respondents to use a randomizing device (e.g., coin flip) and truthfully answer the sensitive question only when the randomization results in a certain outcome. In other cases, respondents simply give a predetermined response.[2] Because enumerators do not know the outcome of randomization, they have no way of knowing whether respondents are answering the sensitive question or providing a preset response.

While each method has strengths and weaknesses as described below, there is little empirical evidence about their relative performance.

2.1 The 2011 Mississippi General Election

An ideal study design to empirically evaluate the performance of these methods would exploit an instance where the ground truth is available for sensitive attitudes or behavior within multiple sub-populations and where there are strong reasons to expect people will not respond truthfully to direct questions. The November 2011 Mississippi General Election provides one such context. This election included a ballot initiative, the so called "personhood amendment" (formally know as Ballot Measure 26), changing the Mississippi constitution to declare that life begins at conception.

Based on interviews and their knowledge of Mississippi politics, most commentators expected the initiative to pass easily. A poll of 796 likely voters taken just 24 hours before the election found with direct questioning that only 44% of likely voters planned to oppose the amendment (Public Policy Polling, 2011). Yet, the amendment was defeated 57.6% to 42.4%, a 14 percentage point swing from the pre-election poll. The difference between the polling results and the final tally was more than 4 times the poll's 3.5 percentage point margin of error and larger than the 11% of undecided voters in the same poll. No similar deviations from the poll were observed elsewhere

---

[2]This is one design of the randomized response method, which Blair *et al.* (2014a) calls the "forced design."

on the ballot. A large portion of Mississippi voters apparently dissembled when asked about this socially sensitive issue but were honest about other items.[3]

Beyond having clear evidence of preference falsification on this issue, two additional facts about Mississippi elections make it an ideal place for the study. First, like other states, the Mississippi Secretary of State makes their voter rolls public, so we can survey people who did in fact vote in the election. Second, although official records do not reveal individual votes, county recapitulation reports provide the true vote share at the precinct level and above, enabling us to assess the empirical performance of various survey methods across multiple political units. This allows us to characterize the bias and variance of each method across units.

2.2 Sample Selection

To maximize the validity of our estimates of voting, we drew a stratified random sample of 2,655 respondents who voted in Mississippi's 2011 statewide general election according to the Mississippi state voter history file. This is a significant improvement over previous studies. For some reason, even research that has focused on voter turnout, a subject for which official data is readily available, has generally failed to capitalize on the opportunity to validate estimates (see Belli *et al.*, 2001, for an exception). Holbrook and Krosnick (2010b), for example, were unable to compare the estimates from their list experiment with the official turnout rates because all of their samples (RDD telephone and non-probability internet sample) include those who did not vote.

To maximize the sensitivity of our question, we drew our sample from 19 counties where support for the referendum was lower than would have been predicted given support for the Republican gubernatorial candidate. This is done by first regressing the proportion of "no" votes on Republican gubernatorial vote share and then choosing the counties with large positive residuals as well as large populations. Figure 1 illustrates this process by plotting the proportion voting "no" on personhood for each county on the vertical axis and the portion voting for the Republican gubernatorial candidate on the horizontal axis. We sampled the 19 counties with the largest positive

---

[3]To the best of our knowledge no new information emerged between the poll and the election that would have suddenly changed the minds of a large number of voters on this issue.
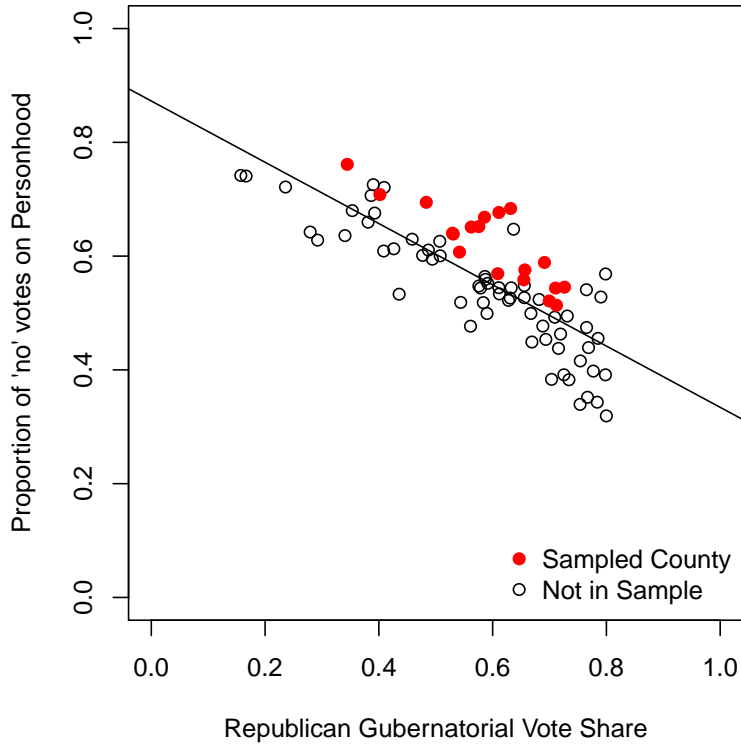
Figure 1: Selection of Counties in the Mississippi Validation Study. This figure shows the strong negative relationship between votes against the Personhood Amendment (the vertical axis) and for the Republican gubernatorial candidate in the November 2011 Mississippi General Election (the horizontal axis) where the solid line represents the linear regression fit. The figure depicts all 82 counties statewide. Red solid circles denote 19 counties selected for the study. These counties are relatively populous and have large positive residuals, implying that they showed an unexpectedly large number of "no" votes on the amendment. The voters of these counties, therefore, may exhibit a large degree of social desirability bias.

residuals that had at least 4,000 voters to allow a sufficient number of potential respondents in each arm of the experiment (as described below). On the whole, the resulting counties are solidly Republican with the median Republican gubernatorial vote share equal to 60.9%. These counties voted unusually strongly against the personhood amendment. The median 'no' vote in our sample counties was 63.9%, substantially larger than the 54.8% state-wide county-level median.

2.3 Survey Instrument

We now describe our survey. The interviews were conducted via phone by a commercial firm, Braun Research, who dialed based on the standard Aristotle national voter file which matches phone numbers to the State of Mississippi's official voter file. Voters whose phone number was

unavailable were treated as unit non-response just like those who did not answer the call or refused to participate in the survey. As explained later, the potential bias due to non-response will be corrected with either weighting or regression adjustments in our analysis. To maximize recall, we began the experiment by reminding respondents of several issues at stake in this election. The opening script reads as follows:

```
Before I ask you any questions, I want to remind you of some of the political
issues in the November 2011 General Election. As you may remember,
Mississippians voted about a number of initiatives to amend the state's
constitution, including the "Voter ID" amendment which required voters to
present ID at the polling station, the "Eminent Domain" amendment which
limited the state's ability to take private property, and an amendment to
declare that life begins at fertilization. In the media, this ballot initiative
was often called the Personhood Initiative. Now we'll ask you some questions.
```

Following this prompt, we employed a nested design. The vast majority of respondents received one or two different indirect methods followed by a direct question and a small portion of respondents received only the direct question. Within each county respondents were randomized into different question orderings. We did not fully randomize across all possible orderings of questions for both practical and other reasons. For example, we never administered indirect questions after the direct question because we found in our pre-test that many respondents refrained from answering indirect questions about politics at higher rates after the direct question was administered.

Direct Question. In our survey, the direct question was administered as follows,

```
Did you vote YES or NO on the Personhood Initiative, which appeared on the
November 2011 Mississippi General Election ballot?

Voted Yes; Voted No; Did not vote; Don't know; Refused
```

The use of direct questioning for sensitive issues has two major advantages. Because the responses are directly observed (so long as people are willing to answer), the direct question is statistically most efficient to analyze and is easy to implement and interpret. Unfortunately, a number of validation studies have found significant underreporting of sensitive behaviors and attitudes with direct questions (e.g., Folsom, 1974; Junger, 1989; van der Heijden *et al.*, 2000; Hessing

et al., 1988; Elffers *et al.*, 1992). Another major disadvantage of the direct questioning is that it often suffers from high item non-response rates. For example, scholars have found that approximately 40% of respondents refuse to answer direct questions about their views towards militant groups in Afghanistan and Pakistan (Blair *et al.*, 2013; Lyall *et al.*, 2013). Such non-response issues are commonly thought to be mitigated when enumerators take the time to build substantial trust and rapport with respondents. However, doing so is costly and is often difficult, and little is known about the efficacy of such strategies.

List Experiment. Under the standard design of the list experiment, researchers randomize a sample of respondents into two groups where a list of several control items is presented to the control group and a list of the same control items plus one sensitive item of interest is read to the treatment group. Respondents are then asked to count the number of items on their list that fit certain criteria rather than asking each item separately. The difference in means between the two groups then provides the simplest estimate of the prevalence of the sensitive attitudes or behavior in a target population though more efficient estimators are now available.

In our study, we used this standard design and asked the following question to the control group,

```
Here is a list of four things that some people have done and some people have not.
Please listen to them and then tell me HOW MANY of them you have done in the
past two years. Do not tell me which you have and have not done. Just tell me
how many:
- Discussed politics with family or friends;
- Cast a ballot for Governor Phil Bryant;
- Paid dues to a union;
- Given money to a Tea Party candidate or organization.
How many of these things have you done in the past two years?
```

For the treatment group, the same exact script was read but the following additional sensitive item of interest was added to the list,

```
- Voted 'YES' on the 'Personhood' Initiative on the November 2011 Mississippi
General Election ballot.
```

Respondents in each group could report the number of items that applied to them and like the direct question they also had "don't know" and "no response" options.

The advantage of this approach is that respondents do not directly report whether the sensitive item applies to them. Instead, they provide a count of items on a list which contains other items. The major limitation of the list experiment is the problem of ceiling and floor effects. Answering "0" to our list experiment in the treatment group, for example, reveals that the person voted "no" on the personhood initiative. Another disadvantage is the fact that the aggregation often decreases statistical efficiency of subsequent analyses.[4] To address these concerns, Glynn (2013) recommends that the the researcher chooses control items such that responses to those items are negatively correlated. Hence, we include an item about paying union dues alongside an item about supporting a Tea Party candidate or organization.

Endorsement Experiment. The endorsement experiment works by exploiting evaluation bias in human judgement. As in the list experiment, a sample of respondents is randomized into two groups. In the control group, respondents are asked to evaluate some relatively uncontroversial issue or object (e.g., rate a policy on a Likert scale). In the treatment group, that issue or object is associated with the sensitive item before being evaluated (e.g., the same policy is said to be endorsed by a controversial political group). The difference between these two groups is then taken to reflect the degree to which respondents are favorable (or unfavorable) towards the sensitive item.

While the endorsement experiment has previously been used to measure attitudes about political figures (e.g., Blair *et al.*, 2013; Lyall *et al.*, 2013), we use it to measure a sensitive political behavior, i.e., voting "no" on the Personhood referendum. To do this, we flip the standard design and ask respondents to rate their support for political actors (which is relatively uncontroversial) and then randomize the pairing of those actors with the support for the sensitive policy of interest. If this pairing induces a negative effect on voters' support level for the political actors, we interpret this effect as evidence for their opposition to the referendum. Specifically, we asked the control

---

[4]Another disadvantage is that adding a sensitive item may alter one's (latent) response to control items (Blair and Imai, 2012).

group,

```
We'd like to get your overall opinion of some people in the news. As I read each
name, please say if you have a very favorable, somewhat favorable, somewhat
unfavorable, or very unfavorable opinion of each person.

Phil Bryant, Governor of Mississippi?

  Very favorable
  Somewhat favorable
  Don't know/no opinion
  Somewhat unfavorable
  Very unfavorable

  Refused
```

And in the treatment group, we added the information that Governor Bryant supported the personhood amendment as follows,

```
Phil Bryant, Governor of Mississippi, who campaigned in favor of the 'Personhood'
Initiative on the 2011 Mississippi General Election ballot?
```

The endorsement experiment is grounded in extensive research on persuasion in social psychology (see Petty and Wegener, 1998, for a review). Researchers have found that individuals are more likely to be persuaded and influenced by likable sources (Petty and Cacioppo, 1986; Cialdini, 1993) and that endorsements of policies and positions are much more effective when an individual has positive affect toward the source of the endorsement (Wood and Kallgren, 1988; Chaiken, 1980; Petty *et al.*, 1983). As O'Keefe (1990) summarizes, "Liked sources should prove more persuasive than disliked sources" (p. 107).

The main advantage of the endorsement experiment is that unlike the list experiment it can never reveal the truthful answer to the sensitive question. However, this indirect nature also presents a major drawback in that a latent variable model is needed to derive estimates of sensitive behaviors from the ordered responses (as discussed in Section 3.2) and the endorsement effects do not have an obvious scale.[5] The endorsement experiment is also statistically inefficient, even when compared with other indirect questioning techniques. For this reason, Bullock *et al.* (2011)

---

[5]They can, however, be benchmarked against the effect for endorsers whose level of support is commonly understood to be strongly positive or negative (e.g. Osama Bin Laden). For an example of this approach, see Fair *et al.*

recommends that the researcher uses multiple questions to measure one sensitive item. The design we use here, with the single item gauging support for Governor Phil Bryant, gives us an estimate of the lower bound of the endorsement experiment's statistical efficiency relative to other approaches.

Randomized Response. The randomized response method obscures individual responses by introducing random noise. A number of designs have been introduced since the work of Warner (1965) (see Blair *et al.*, 2014a). In our Mississippi study, we adopt the standard forced response design where a coin flip is used for randomization. Because the randomized response is thought to be difficult for respondents to grasp, we first gave respondents a chance to practice by asking about whether they voted. We then proceeded to ask the main question about their vote on the Personhood amendment. Our script is given here,

```
To answer this question, you will need a coin. Once you have found one, please
toss the coin two times and note the results of those tosses (heads or tails) one
after the other on a sheet of paper.  Do not reveal to me whether your coin lands
on heads or tails.  After you have recorded the results of your two coin tosses,
just tell me you are ready and we will begin.

First, we will practice. To ensure that your answer is confidential and known
only to you, please answer 'yes' if either your first coin toss came up heads or
you voted in the November 2011 Mississippi General Election, otherwise
answer 'no'.

    Yes
    No
    Don't know
    Refused

Now, please answer 'yes' if either your second coin toss came up heads or
you voted 'YES' on the 'Personhood' Initiative, which appeared on the
November 2011 Mississippi General Election ballot.

    Yes
    No
    Don't know
    Refused
```

To discourage break-offs, interviewers were instructed to use the following script if respondents

(2014) who compare endorsement effects for various militant groups in Pakistan to those for Abdul Sattar Edhi, a widely-revered philanthropist.

expressed confusion or hesitation about the instruction to find a coin:

```
I mentioned earlier that this research is to help us better understand how to
ask people about political issues. It may seem strange that we asked you
to find a coin. Sometimes survey respondents want to keep their answers
to questions private. We're going to ask you some questions in a way that
lets you keep your answers secret, even from me.    But, we need a coin to
make it work. All of this will be clear in second.
```

In pre-testing, this script helped to reduce nonresponse on the randomized response items.

Because we were only surveying people who had voted according to the voter file, the first practice question provides another check on whether the randomized response is working. We find that about 90% of those who answered the question (10% refused) gave the correct "Yes" answer. It is unclear why approximately 8% of respondents answered "No" especially when the socially desirable answer here is "Yes." One possibility is that they are confused about the procedure, which is one of the weakness of the randomized response method identified by the literature.

A main disadvantage of randomized response is the burden it imposes on respondents. The method requires respondents to administer randomization on their own, and this can lead to a high rate of refusal and attrition. Both Coutts and Jann (2011) and Holbrook and Krosnick (2010a) flag major problems involving respondents' noncompliance with the randomized response instructions. Indeed, these authors find that randomized response produces more nonresponse and less valid estimates than a list experiment.

However, these studies do not compare randomized response estimates against the truth and, as a result, their conclusion may not be entirely warranted. Indeed, according to a comprehensive review article by Lensvelt-Mulders *et al.* (2005), several studies, which validate estimates against the true sensitive information, find that the randomized response technique performs reasonably well. The present study offers the first validated comparison of the randomized response technique against other indirect questioning methods. To preview the results, we find that randomized response produces estimates which compare favorably with those of both the direct question and other indirect techniques.

3 Statistical Analysis of Indirect Questioning Methods

In this section, we describe the statistical methods we use to analyze the responses from indirect questioning techniques. Many of the methods used in this paper are explained in more detail elsewhere and hence interested readers should consult these other papers (Bullock *et al.*, 2011; Imai, 2011; Blair and Imai, 2012; Blair *et al.*, 2014a).

3.1 List Experiment

Suppose that we have a random sample of n survey respondents from a target population. Let J represent the total number of control items on a list. As explained above, under the most basic design of the list experiment, we randomly divide the sample into two groups. In the control group, respondents are asked to report the number of items from the list of J control items they answer affirmatively. In the treatment group, on the other hand, the respondents are exposed to the total of $J + 1$ items which includes an additional sensitive item of interest as well as the same set of J control items. We use Y_i to denote the observed response for each respondent and $T_i = 1$ $(T_i = 0)$ to represent that respondent i is assigned to the treatment (control) group.

Imai (2011) and Blair and Imai (2012) show how to conduct a multivariate regression analysis using the responses from list experiments. In our analysis, we use the logistic regression to model the latent response Z_i^* to the sensitive item given a vector of respondent demographic characteristics X_i obtained from the voter file.

$$\Pr(Z_i^* = 1 \mid X_i) \quad = \quad \text{logit}^{-1}(\alpha + \beta^\top X_i) \tag{1}$$

where (α, β) is a vector of coefficients. The model is completed with the following binomial sub-model for the responses to the control items Y_i^*,

$$\Pr(Y_i^* = y \mid X_i, Z_i^*) \quad = \quad \binom{J}{y} g(X_i, Z_i^*)^y \{1 - g(X_i, Z_i^*)\}^{J-y} \tag{2}$$

where $g(X_i, Z_i^*) = \text{logit}^{-1}(\gamma + \delta^\top X_i + \zeta Z_i^*)$ and (γ, δ, ζ) is a set of coefficients. These latent variables are related to the observed response Y_i via the relationship, $Y_i = T_i Z_i^* + Y_i^*$. Imai (2011) and Blair and Imai (2012) show how to obtain the maximum likelihood estimates of these parameters via the

EM algorithm.

For the county level analysis described later in this paper, we will use the random intercept model where each county is allowed to have a different intercept. Specifically, the above model is modified as $\Pr(Z_i^* \mid X_i) = \text{logit}^{-1}(\alpha_{\text{county}[i]} + \beta^\top X_i)$ and $g(X_i, Z_i^*) = \text{logit}^{-1}(\gamma_{\text{county}[i]} + \delta^\top X_i + \zeta Z_i^*)$ where both $\alpha_{\text{county}[i]}$ and $\gamma_{\text{county}[i]}$ follow a normal distribution. For the estimation of this model, we use a Bayesian framework with an uninformative prior and employ a Markov chain Monte Carlo algorithm that is similar to the one developed by Blair $et\ al.$ (Forthcoming). Both of these models are implemented in the R package `list` (Blair and Imai, 2013) and we use this package for our subsequent analysis.

3.2 Endorsement Experiment

We utilize the statistical model proposed by Bullock $et\ al.$ (2011) to analyze the data from the endorsement experiment. Under the standard design, we randomly split the sample of n respondents into two groups. In a typical endorsement experiment, for the respondents in the control group, a policy is described and they are asked to rate their level of support for the policy. In the treatment group, the respondents are asked to do the same but the policy is said to be endorsed by an actor. If this endorsement increases the level of support for the policy, then we interpret this effect as evidence that a respondent holds a favorable view towards the actor.

As explained above, in our Mississippi endorsement experiment, we are interested in estimating voting on the sensitive abortion referendum. This is done by asking respondents to rate their support for a politician while for those in the treatment group we mention the fact that the politician supported the Personhood referendum as the "endorsement." If this additional piece of information decreases respondents' support for the politician, we interpret this effect as evidence that they oppose the referendum.

Let Y_i represent the K category ordered response (i.e., $Y_i \in \{0, 1, \ldots, J-1\}$) indicating the reported support level for respondent i towards and T_i be the treatment indicator. We use the following single item ordered probit model, which is a special case of the item response theory

15

based model proposed by Bullock *et al.* (2011),

$$Y_i^* \overset{\text{indep.}}{\sim} \mathcal{N}(\beta(x_i^* + T_i s_i^*) - \alpha, \ 1) \tag{3}$$

where Y_i^* is a latent outcome variable and $Y_i = y$ if $\tau_y < Y_i^* < \tau_{y+1}$ with the cut-points $\tau_0 = -\infty, \tau_1 = 0$, and $\tau_K = \infty$. The latent variale x_i^* is the ideological position of respondent i and s_i^* is the additional support level induced by the endorsement.

Following Bullock *et al.* (2011), we model x_i^* and s_i^* hierarchically as follows,

$$x_i^* \overset{\text{indep.}}{\sim} \mathcal{N}(\delta^\top X_i, 1) \tag{4}$$

$$s_i^* \overset{\text{indep.}}{\sim} \mathcal{N}(\lambda^\top X_i, \omega^2) \tag{5}$$

The model is completed by specifying uninformative prior distributions on model parameters $(\alpha, \beta, \delta, \lambda, \omega^2)$. As shown by Bullock *et al.* (2011), this model can be easily extended to a random intercept model, which we use for our county-level analysis. To fit these models, we use the R package `endorse` (Shiraito and Imai, 2012) and implement a Markov chain Monte Carlo algorithm.

To link this model directly to the model for the list experiment, Blair *et al.* (Forthcoming) suggest that researchers use the posterior probability of having a positive endorsement effect, i.e., $\Pr(s_i^* > 0)$, as the main quantity of interest and interpret it as the estimated probability of positive support for the sensitive actor or policy. We will use this as the estimated probability of vote choice on the Personhood referendum. We adopt this approach when analyzing our Mississippi experiment.

3.3 Randomized Response

While methodological work on list and endorsement experiments has been relatively rare until recently, researchers have developed various statistical methods for analyzing the data from the randomized response method. This literature goes back to Warner (1971), who formulated a general linear model. Recent work has extended this approach to non-linear models such as logistic regression (e.g., van den Hout *et al.*, 2007).

Under the standard forced response design of randomized response utilized in the Mississippi study, respondents are asked to flip a coin in private and truthfully answer the sensitive question

on the Personhood referendum if the coin lands on heads. If the coin lands on tails, however, they are asked to answer "no," which is the sensitive answer in this case. Let Y_i represent the observed response whereas we use Z_i^* to denote the latent truthful answer to the sensitive question. We use the logistic regression model for the latent response, which is identical to the model for the list experiment given in equation (1). Following van den Hout *et al.* (2007), the likelihood function for this model under the randomized response method is given by,

$$\prod_{i=1}^{n} \left\{ p \cdot \text{logit}^{-1}(\alpha + \beta^\top X_i) \right\}^{Y_i} \left\{ 1 - p \cdot \text{logit}^{-1}(\alpha + \beta^\top X_i) \right\}^{1-Y_i} \tag{6}$$

where $p = 1/2$ is the probability of a coin landing on head.

Blair *et al.* (2014a) derive the EM algorithm to reliably estimate this and other randomized response methods, and we follow their approach. For county-level analysis, we again use a random intercept model as in the case of the list experiment. To fit this random intercept model, we adopt the Bayesian approach of Blair *et al.* (2014a) with a non-informative prior and employ their Markov chain Monte Carlo algorithm. We use the R package `rr` (Blair *et al.*, 2014a) for all of our computation.

4 Empirical Findings

In this section, we describe our empirical findings. We begin by examining the bias of the direct question and then compare it with the performance of randomized response, list experiment, and endorsement experiment. In our comparison, we will use three common ways of adjusting for unit and item non-responses by computing unweighted, weighted, and regression adjusted estimates.

4.1 Bias of Direct Question

We first investigate the performance of the direct question. A number of previous studies have found that when asking sensitive survey questions many respondents exhibit social desirability bias by concealing socially undesirable attitudes and behavior while others choose the "don't know" category or refuse to answer the question (see e.g., Tourangeau and Yan, 2007). This is exactly what we find in our study.

In the left two columns of Table 1, we observe that only 30% of survey respondents admit

Direct Question		List Experiment			Endorsement Experiment			Randomized Response	
			Treatment	Control		Treatment	Control		
voted yes	0.454	have done 0	0.099	0.149	very favorable	0.315	0.314	yes	0.594
voted no	0.305	1	0.231	0.306	somewhat favorable	0.288	0.305	no	0.274
did not vote	0.046	2	0.276	0.397	indifferent	0.093	0.085		
		3	0.252	0.114	somewhat unfavorable	0.137	0.122		
		4	0.066	0.020	very unfavorable	0.126	0.104		
		5	0.050						
non-response	0.194	non-response	0.026	0.015	non-response	0.040	0.069	non-response	0.133
Mean	0.598		2.106	1.544		2.464	2.353		0.685
Sample size	2655		666	686		902	939		943

Table 1: Summary of Survey Responses to the Direct Question, the Randomized Response, the List Experiment, and the Endorsement Experiment. The table presents the share of respondents in each answer category, separately for the treatment and control groups when appropriate. The last row shows the number of respondents randomly assigned to each question and condition. Proportions do not sum to 1 due to rounding. Non-response represents that respondents either chose the "don't know" category or refused to answer the question.
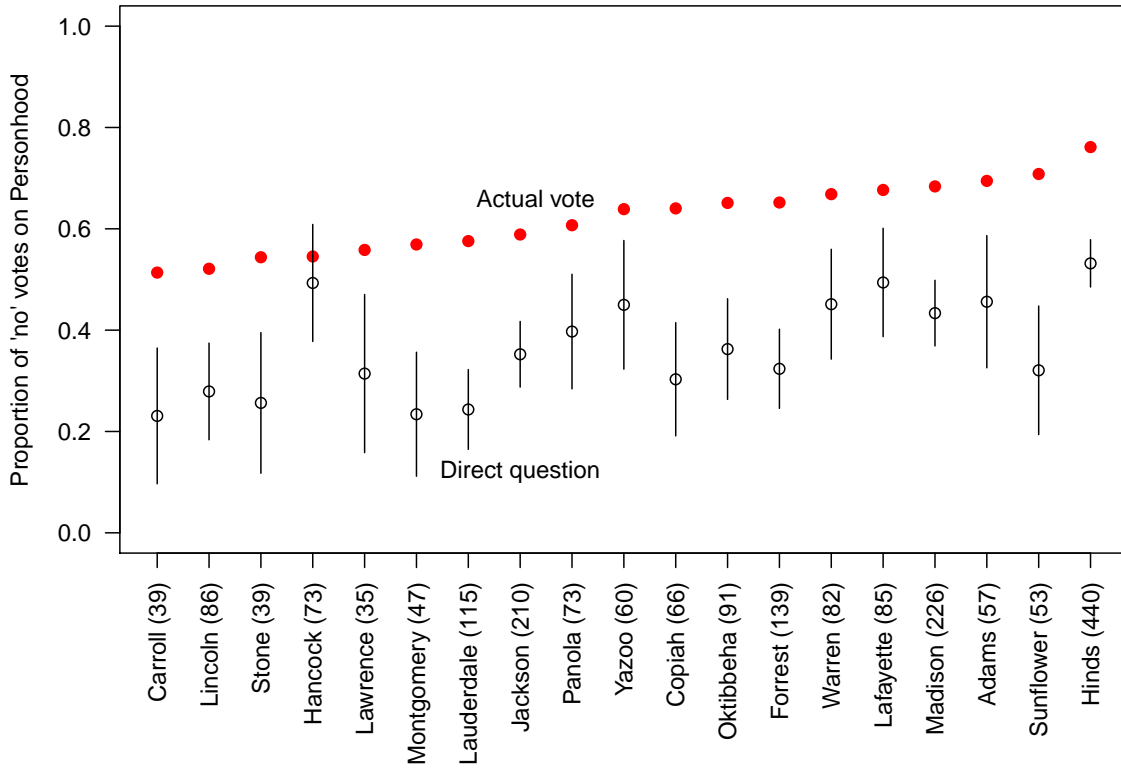
Figure 2: Social Desirability Bias of the Direct Question. The figure compares the county-level estimates of the direct question (open circles) with the actual "no" votes for Personhood (solid red circles). The vertical lines represent 95% confidence intervals. Across all counties but one, the direct question severely underestimates the actual vote share by 20 percentage points and more. The sample size for the direct question in each county is given in parentheses.

voting "no" on the referendum, a socially undesirable act in this context, while the official election record shows 65.3% actually voted no across the 19 counties in our study. In fact, as in the pre-election surveys mentioned above, the direct question suggests that support outweighed opposition to Personhood by 15 percentage points and that the referendum would have passed. Moreover, almost 20% of respondents either said "don't know" (17.2%) or refused to answer the question (2.3%), reflecting the question's highly sensitive nature.

Figure 2 shows the nature of social desirability bias by displaying the county-level estimates based on the direct question (open circles) together with the actual "no" vote shares (red solid circles) together in a single plot. The vertical lines represent 95% confidence intervals and counties

are ordered according to the actual "no" vote share on the horizontal axis. Therefore, the differences between the estimates and the actual "no" vote share represent the social desirability bias of the direct questioning. We observe that across all counties the direct question severely underestimates the actual vote share. The magnitude of bias is large, often exceeding 20 percentage points. As shown below, these results remain largely identical under different weighting schemes.

4.2 Pooled Analysis

We next investigate the performance of indirect questioning methods as well as that of the direct question for the entire sample under different weighting schemes.

We begin by examining how willing respondents were to answer using each method. As Table 1 shows the non-response rates for all three indirect methods are significantly lower than for the direct question. Though previous studies of randomized response have documented higher non-response rates stemming from the technique's complexity (Buchman and Tracy, 1982; Coutts and Jann, 2011),[6] we find that respondents were more willing to answer the randomized response question than the direct question by more than 5 percentage points. All of these non-responses are due to "refusal," presumably due to the complex nature of the question which imposes higher cognitive demands for respondents. The list and endorsement experiments even have much lower non-response rates than the randomized response, approximately 2% and 5%, respectively. This finding is consistent with the expectation that respondents often find the list and endorsement experiment questions less obtrusive and easy to understand and answer than the randomized response.[7]

The poor performance of the direct question we identified above is due to two sources of bias, namely non-response and social desirability. We employ three commonly used strategies to adjust for non-response bias. First, we estimate the proportion of voting "no" on Personhood by listwise deleting non-responses. We call these estimates *unweighted*. Second, we calculate survey weights for respondents using the demographic information from the voter file and compute the *weighted* average of the indicator variable for reporting "no" vote on the referendum. Specifically, we calcu-

---

[6]But see also Wolter and Preisendorfer (2013) and Lara *et al.* (2004) who found equivalent response rates for randomized response.

[7]A fact also consistent with our intensive pre-testing of the questionnaire.
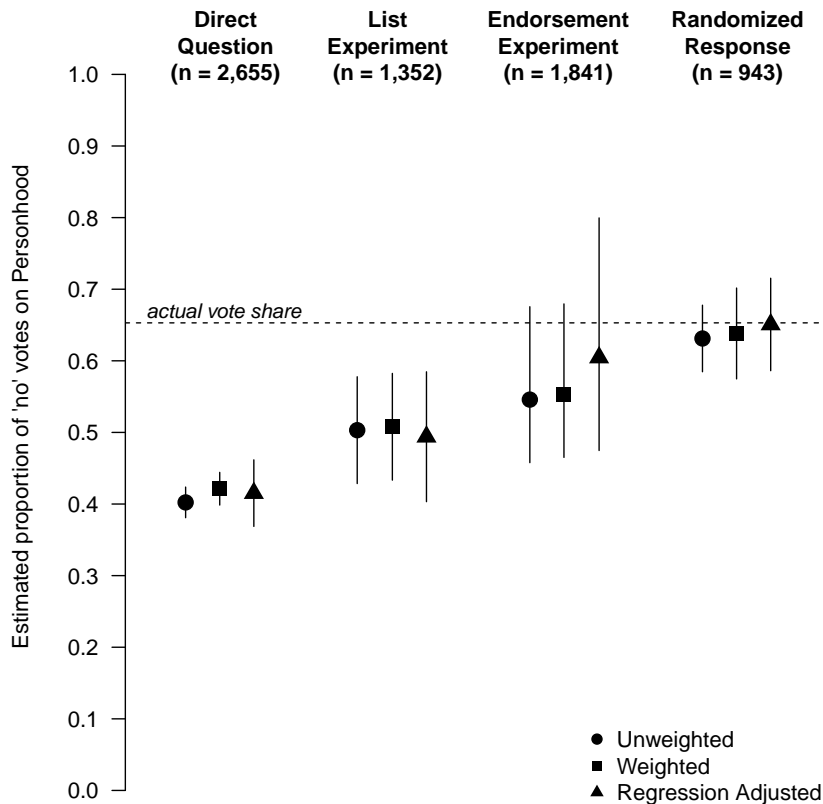
Figure 3: Comparison of the Direct Question with the Three Indirect Methods and the Actual Vote Share. This figure compares the estimated proportion of the sensitive behavior, voting against the Personhood referendum, using the direct question, the list experiment, the endorsement experiment, and the randomized response technique. For each method, the figure presents three types of estimates: unweighted (circles), weighted (square), and regression adjusted (triangles). The actual vote share is represented by the dotted line, and the vertical bars indicate 95% confidence intervals.

late weights by regressing age,[8] gender, party ID, and county on the probability of inclusion in each experimental condition in a binomial logistic regression.[9] We use these regression-based weights rather than stratification-based weights due to the sparse nature of the demographic information available in the voter file. Finally, we also adjust for the lack of representativeness of the sample

---

[8]Age is included in the regression as a categorical variable with six levels: 25-34, 35-44, 45-54, 55-64, 65plus, and missing age.

[9]Specifically, we fit a bayesian binomial logistic regression using the arm package with default non-informative priors (Gelman *et al.*, 2008). The probability of inclusion in the sample was then calculated based on the fitted coefficients and the weights were defined as the inverse of this probability. Weights were then trimmed to a maximum value of 20 times the smallest weight.

via regression. We aggregate predicted vote choice for all individuals who the official records indicate have turned out in this election. The predictions are obtained from the logistic regression of self-reported vote choice using the aforementioned covariates. We call these estimates *regression adjusted*.

Figure 3 shows the resulting three estimates with 95% confidence intervals. For the direct question (the left most estimates), we observe that each adjustment makes little difference to the bias of the original estimate (listwise deletion indicated by the solid circle while weighting and regression adjustments represented by the solid square and triangles). Indeed, these estimates are still severely biased. Of course, this may be in part because Mississippi's voter file lacks detailed demographic information. For example, date of birth is missing for most voters. Nevertheless, the results clearly show that the measure based on the direct question is unreliable.

The same plot also presents the estimates from indirect questioning methods. For these methods, weighting adjustment is conducted by first fitting the statistical models described in Section 3 without covariates and then obtaining the weighted average of posterior predicted values of vote choice across survey respondents. For regression adjustment, we use the same models with covariates to obtain posterior predictions of vote choice for all voters in the voter file and compute their (unweighted) average. Similar to the direct question, the three estimates (unweighted, weighted, and regression adjusted) for the indirect methods are also statistically indistinguishable from each other; however, these estimates are much closer to the actual vote share. In particular, the estimates based on the endorsement and, especially, the randomized response perform well. While they still underestimate the vote share somewhat, these estimates are no longer statistically distinguishable from the actual vote share.

4.3 County-level Analysis

One major advantage of our study design is that we can validate the estimates against the actual election outcomes at the county-level. This allows us to quantify how these methods perform on average across the 19 counties. Figure 4 reports the same set of three estimates for each method: unweighted, weighted, and regression-adjusted. The models fitted in this county-level analysis
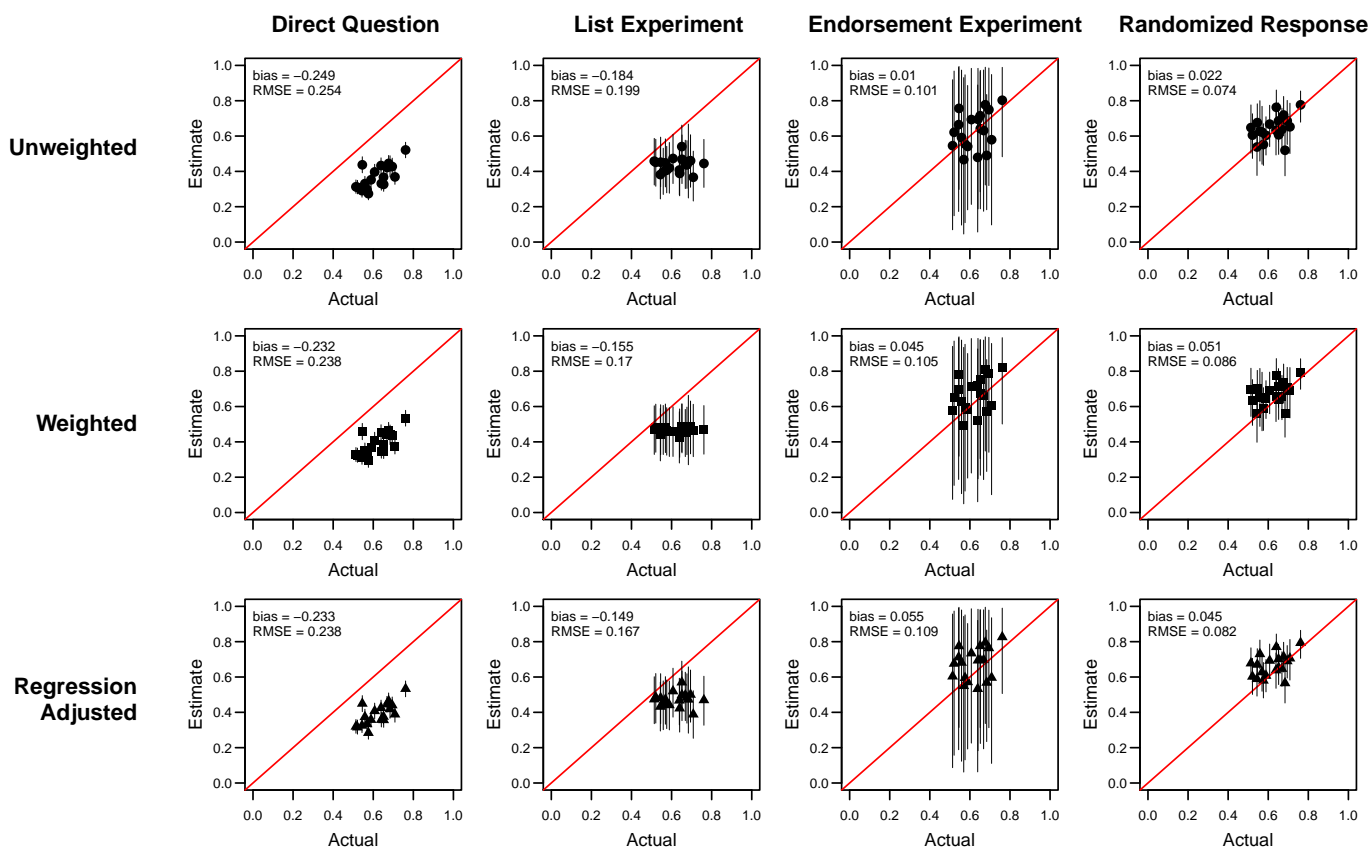
Figure 4: County-level Comparison of the Estimates based on the Direct Question with Those based on the Three Indirect Methods. The plots compare county-level estimates of the sensitive behavior, voting against the Personhood referendum, for all questioning methods and estimation approaches. County-level estimates (y-axis) are plotted with the 95% confidence intervals against actual vote share (x-axis) with points below the 45 degree line indicating the underestimation. The bias and root mean squared error (RMSE) across counties are also presented.

include random county-level intercepts to account for heterogeneity across counties. For the direct question, we fit the logistic regression with random intercepts, which is also an underlying model for the three indirect questioning methods under consideration. In each plot of the figure, we directly compare the county-level estimates (with 95% confidence intervals) on the vertical axis against the corresponding actual vote share on the horizontal axis. Points below the 45 degree lines, therefore, represent under-estimates.

The results further suggest that indirect questioning methods reduce bias relative to the direct questions. For any given method, the magnitude of bias across counties is greatest for direct questioning, while the endorsement experiment and the randomized response exhibit the least

23

amount of bias. In particular, the performance of the randomized response is impressive as its estimates are much less noisy than those of the endorsement experiment. The list experiment hits the middle ground between the direct question and the other two methods. It is less biased than the direct question but the magnitude of its bias is much greater than the endorsement experiment and the randomized response.

In addition, the variance of the list experiment estimates is smaller than that of the endorsement experiment, but the list estimates are less precise than those based on the randomized response. Unlike the same model applied to the other two indirect questioning techniques, the random intercept model for the list experiment recovers poorly the trend of actual election results across counties. While the difference-in-means estimator of the list experiment has a reasonable positive correlation with the actual election results (0.382 with p-value of 0.107), many of these simple estimates exceed the logical range of $[0, 1]$. When applying the random intercept model, however, the county-level data are too noisy and the model essentially yields the pooled estimate for all counties.

In sum, the results show that while the direct question is severely biased this bias can be reduced by the use of indirect questioning techniques. Among these survey methods, the randomized response recovers the truth well. The endorsement experiment is less biased than the list experiment but is noisier.

Finally, we aggregate the county-level results given in Figure 4 across the counties to obtain overall estimates that are comparable with the estimates based on the pooled analysis given in Figure 3. Figure 5 shows these results. For all methods except for the endorsement experiment these estimates are similar to the corresponding estimates given in Figure 3. For the endorsement experiment, aggregating from county-level estimates appears to reduce bias to some extent, suggesting that the α and β coefficients in equation (3) vary significantly by county and taking account of this heterogeneity improves the estimates.

Broadly speaking, the fact that sub-unit estimates, when aggregated, match the pooled individual-level analysis gives us greater confidence that the modeling assumptions at the individual level are
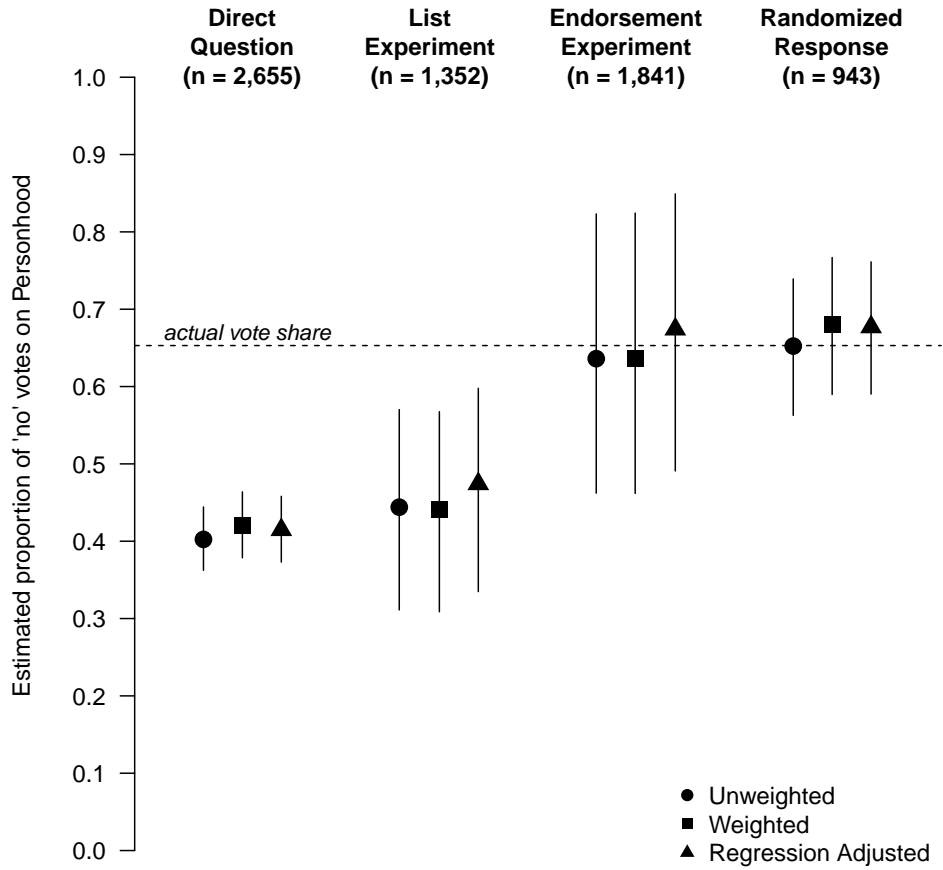
Figure 5: Comparison of the Direct Question with the Three Indirect Methods and the Actual Vote Share based on the Aggregation of County-level Estimates. This figure compares the estimated proportion of the sensitive behavior, voting against the Personhood referendum, using the direct question, the list experiment, the endorsement experiment, and the randomized response technique. Unlike Figure 3, which is based on the pooled analysis, the results in this figure are based on the aggregation of the county-level estimates from Figure 4. For each method, the figure presents three types of estimates: unweighted (circles), weighted (square), and regression adjusted (triangles). The actual vote share is represented by the dotted line, and the vertical bars indicate 95% confidence intervals.

reasonable. Aggregating in this manner also confirms the main finding that indirect methods have significantly less bias than the direct question though they have a greater variability.

4.4 Diagnostics for List Experiment and Randomized Response

Before we present the results of our efficiency and individual-level analyses, we perform diagnostic analyses for list experiment and randomized response. First, the standard analysis of list exper-

iments, including the model used, here assumes no design effect (i.e., the addition of a sensitive item does not affect respondents' answers to the control list) and no liars (i.e., respondents do not lie about the sensitive item). Blair and Imai (2012) develop a statistical test for detecting the violations of these assumptions.

The application of this test to the list experiment in our Mississippi survey suggests that these identifying assumptions appear to be violated (Bonferroni corrected p-value of this joint test is 0.003). In particular, we obtain a statistically significant negative estimate (-0.03 with the standard error of 0.01) for the proportion of those who would truthfully answer "No" to the personhood question and "4" to the control item list question. It is unclear as to why this apparent violation arose. Neither the ceiling nor floor effects can explain it. However, the relatively poor performance of the list experiment compared to other indirect methods may at least in part reflect this design effect problem.

Second, the standard forced response design adapted here for randomized response assumes that respondents properly flipped a fair coin and followed the instruction. To probe this assumption, our survey included a question that asks respondents the outcome of the first coin flip for the practice question about turnout.[10] The proportion of respondents who answered "heads" to this question is 56% (65% of those who did not refuse the question), a significant deviation from the expected proportion of 50%. This suggests that the assumption of the standard design and analysis may have been violated.

The magnitude of the deviation appears to depend on age and education. For example, 60.7% of those with at least a college education reported "heads" on the coin toss, compared with 67.8% of those with less education (*p*-value of 0.041). And while 65.6% of respondents age 35 and older reported that their coin landed on "heads," only 52.9% of those under 35 did the same (*p*-value of 0.067). In sum, apparent problems with the randomization were somewhat worse among older and less educated respondents, but deviations from the expected distribution were not unique to these

---

[10]This question reads: Since I do not know the outcome of your coin toss, your answer did not reveal to me whether you voted in the November 2011 Mississippi General Election. To check that you understand our method, would you please tell me whether your first coin toss was heads or tails?

groups.

There are at least two potential explanations for this discrepancy. First, some respondents may have felt uncomfortable about the turnout question and dishonestly reported the outcome of their coin flip. This could explain the upward bias of the coin flip outcome, because the social desirability bias for the turnout question is known to be positive. Second, some respondents may not have actually flipped a coin and simply answered "heads." This explanation assumes that the satisficing answer in this case is "heads."

Under either scenario, however, it is unclear how the estimates for the second question about the abortion referendum would be affected. First, respondents may have correctly implemented the randomization procedure but misreported the outcome of the coin toss due to the sensitivity of the turnout question. If this is the case, the design assumptions for the abortion question may not have been violated. Second, we can test the sensitivity of our results to the alternative assumption that our respondents come from a mixture of those who actually flip a fair coin and those who just satisfice by answering "heads." If we further assume that these two types of respondents have the same probability of answering "yes" to the sensitive question, we estimate the probability of voting against the Personhood amendment to be 54.9% (with a 95% confidence interval of [51.6, 58.3]).[11] This analysis suggests that even in the case where the original design assumptions are violated, the randomized response may still be less biased than direct questioning.

4.5 Efficiency Comparison

The use of indirect questioning methods, while reducing bias, typically results in an efficiency loss over direct questioning. In this section, we compare the efficiency of the direct question with the three indirect methods in order to help researchers better understand the tradeoffs they face when choosing among these techniques.

To facilitate a direct comparison among methods, we sampled an equal number of respondents from each condition. In total, 940 respondents were drawn from those who were assigned to answer each type of question. Specifically, the sampling procedure consists of the following two steps. We

---

[11]The weighted and regression adjusted estimates are: 55.8% [50.3, 61.2] and 56.9% [51.0, 62.9] respectively.

begin by randomly sampling 540 respondents from among those who got each method first based on our nested design. We then drew another 400 respondents from those who got the method second or third. In the case of the direct question, we took all 360 respondents who were first asked the direct question and then drew the remaining 580 from those who got the direct question second or third. Stratifying the random sample of respondents in this way, by question order, accounts for the possibility that respondents who answer using multiple methods differ systematically from those who answer only one question or are answering for the first time.

Figure 6 shows the comparison among methods for the equivalently sized samples. Both unweighted and weighted estimates are given based on models with an intercept only. While the direct question is clearly the most efficient, randomized response also yields shorter confidence intervals than the other two indirect questioning methods. The results for both list and endorsement are significantly noisier. For endorsement in particular, the variance is more than ten times as great as for randomized response. Note that the endorsement experiment in this study used only a single item. This illustrates the need to use multiple items in endorsement experiments for improved efficiency as recommended by Bullock *et al.* (2011).

4.6   Individual-Level Analysis

Finally, we conduct an individual-level analysis. Ideally, we would like to identify the subpopulation for which each of the methods works best. Unfortunately, the true vote shares among subpopulations of individuals (other than those in the same administrative units such as counties and precincts) on the Personhood referendum are unknown. Therefore, we simply compare the pattern of responses across methods for different subgroups. Specifically, we investigate whether the estimates based on indirect questioning methods differ from that of direct questioning for each subgroup.

Our analysis focuses on predicted support for the sensitive referendum by gender, party identification, and educational level. We begin by conducting a multivariate regression analysis (based on the statistical methods described in Section 3) using the survey-measured covariates for age, age squared, gender, party, and education. We use survey-measured variables in this analysis, rather
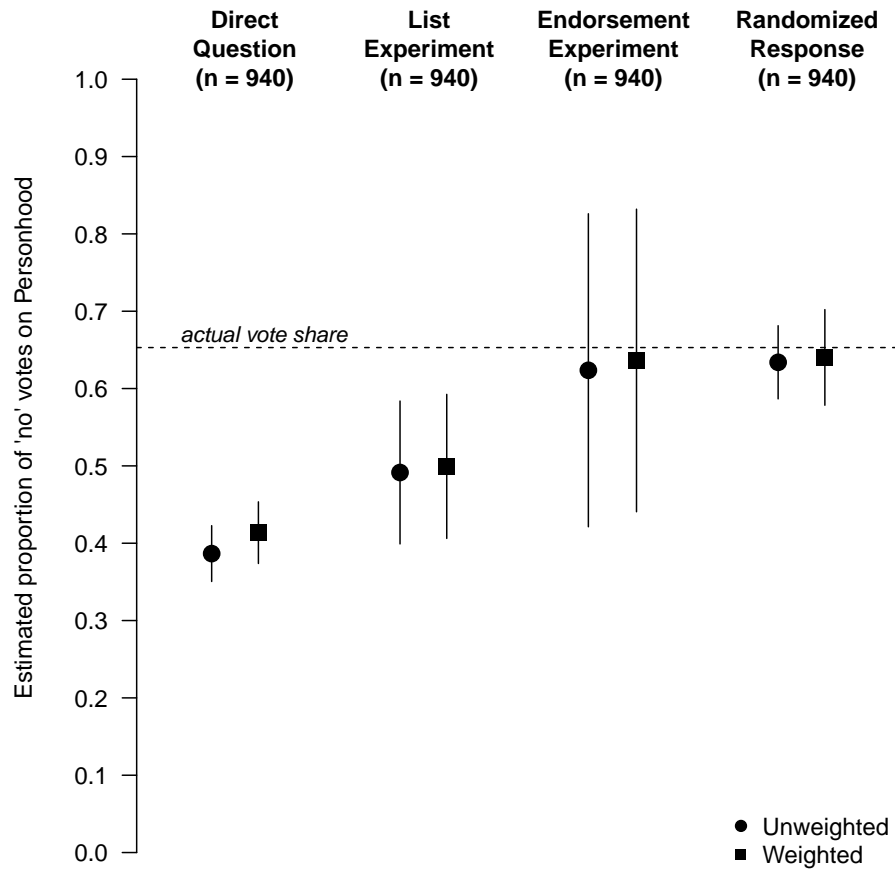
Figure 6: Efficiency Comparison of the Direct Question with the Three Indirect Methods. This figure compares the estimated proportion of the sensitive behavior, voting against the Personhood referendum, using the direct question, the list experiment, the endorsement experiment, and the randomized response technique. However, in contrast to Figure 3, it does so for samples of identical size in order to facilitate a comparison of the efficiency of each method. While the direct question is most efficient, randomized response also fares well. List and, especially, endorsement, are less efficient. Unweighted estimates are again given by circles and weighted estimates by squares, both based on models with an intercept only. The actual vote share is represented by the dotted line, and the vertical bars indicate 95% confidence intervals.

than the voter file covariates used in the other analyses, due to the high level of missingness and limited scope of available variables from the Mississippi Secretary of State. Results for age are not shown here, as neither age nor age squared was statistically significant in the models.

Figure 7 presents the results and suggests that preference falsification may exist among all of the groups examined. The estimates based on the indirect methods for each subgroup are substantially higher than that of the direct question, consistent with underreporting of the sensitive item across
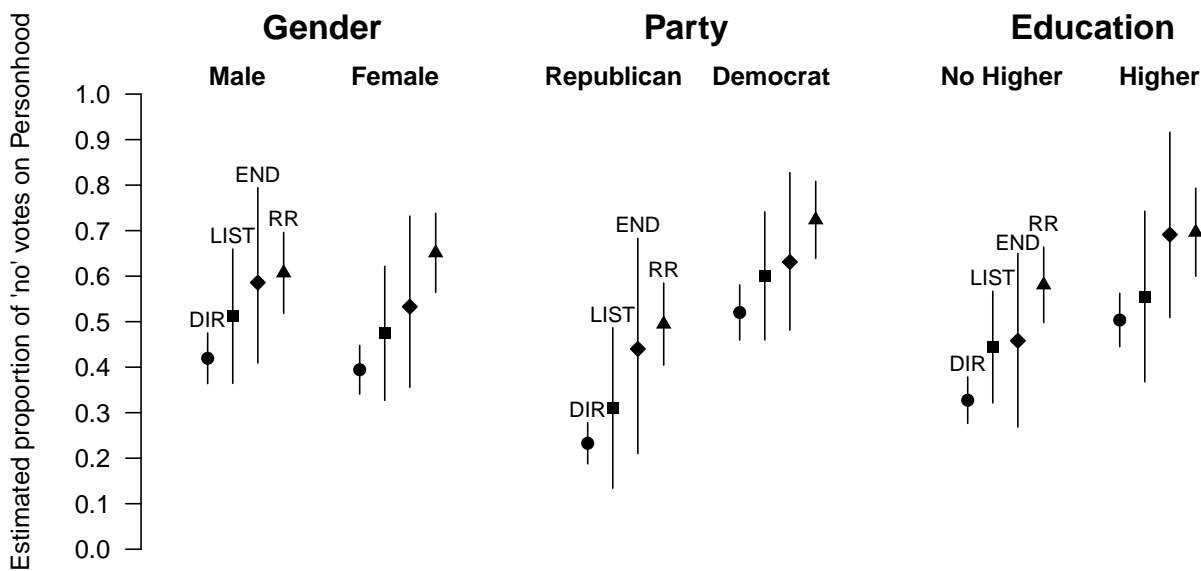
Figure 7: Comparison of Responses Across Subgroups Based on Models with Individual-level Co-variates. This figure compares the estimated proportion of the sensitive behavior, voting against the Personhood referendum, across several categories of respondents based on gender, party identification, and educational level. The results in this figure are based on survey-measured covariates. For each subgroup, the figure presents four estimates using: the direct question (DIR), the list experiment (LIST), the endorsement experiment (END), and the randomized response technique (RR). These results show a consistent pattern of responses for each method, suggesting that preference falsification is present among all of the groups examined. The vertical bars indicate 95% confidence intervals.

the population in response to direct questioning. This analysis further suggests that the indirect methods may have produced results, which are closer to the truth, by reducing social desirability bias across many types of individuals rather than by improving the quality of estimates for some particular group, e.g. Republicans or women. Since the lack of statistical power prevents us from reaching a definitive conclusion, we leave the important question of heterogeneity of social desirability bias to future research.

5   Concluding Remarks

This paper reports the results of the first comprehensive validation study of commonly used survey methods for eliciting truthful responses to sensitive questions. Specifically, we examine the performance of four methods: direct questioning, the list experiment, the endorsement experiment, and the randomized response technique. As these methods become popular among social scientists, it

	Direct Questioning	List Experiment	Endorsement Experiment	Randomized Response
Non-response	most	minimal	minimal	some
Bias	most	some	minimal	least
Variance	least	some	most	minimal
Privacy	none	some	most	most
Cognitive difficulty	least	some	minimal	most

Table 2: The Comparison of Four Common Survey Methods for Eliciting Truthful Responses to Sensitive Questions. This table summarizes our empirical findings and highlights the tradeoffs researchers face in choosing between these survey methods.

is important to learn lessons about when they do and do not work. The best way to do this is to empirically validate findings from these methods against the true sensitive information. We have exploited a unique opportunity that arose in the 2011 Mississippi general election where we knew ground truth, had strong reasons to suspect that direct questioning would perform poorly, and could sample those who participated in the sensitive referendum with near certainty.

Table 2 summarizes our empirical findings and highlights the tradeoffs researchers face in choosing among four survey methods for eliciting responses to sensitive questions. Our core finding is that indirect methods dramatically reduce both non-response and social desirability biases. Non-response on how people voted on the personhood amendment goes from about 20% on the direct question, to 13% on the randomized response, to 6% on the endorsement experiment, to 2% on the list experiment. The bias in our weighted estimates of county-level support for the referendum drops from 0.215 in the direct question, to 0.144 in the list experiment, to 0.042 in the randomized response, to 0.030 in the endorsement experiment. The difference between our list and endorsement experiment results also seem to support the contention of Blair *et al.* (Forthcoming) that list experiments are more prone to social desirability bias than endorsement experiments (p. 29).

Of particular note is the fact that the randomized response performs quite well. This is surprising given the criticism it has received in some recent studies and the evidence that up to 8% of our randomized response respondents may also have been confused about the procedure. We believe that randomized response methods deserve more attention from applied researchers. While this study has employed the most frequently used design of the randomized response technique, there

are other designs that can improve upon the standard forced response (Blair *et al.*, 2014a). In particular, some designs do not require researchers to assume that the randomization distribution is known and provide a greater degree of privacy protection to respondents. Future research should validate these alternative designs.

Future research about eliciting sensitive attitudes and behaviors could take two important directions. First, it remains an open question whether the methods studied here reduce experimenter demand effects in the context of randomized trials. As Bursztyn *et al.* (2014) show, experimenter demand effects can persist even with anonymized behavioral measures of political attitudes. Whether and how much such effects are ameliorated through indirect questioning is an important topic for future research.

Second, more methodological work is needed to evaluate the bias-variance tradeoff concerning the direct and indirect questioning techniques. Such work will help applied researchers make informed methodological choices by properly calibrating the consequences of bias, value of statistical precision, and cost constraints in each particular setting. While we have shown that indirect questioning methods may reduce response bias, they produce estimates that are less efficient than direct questioning. Recent research has sought to address this issue by developing statistical models which combine multiple experimental techniques, such as list and endorsement, in order to recoup this loss of efficiency (Blair *et al.*, Forthcoming). A natural extension of this line of research is to incorporate randomized response into a common statistical model.

An important step in both agendas will be finding opportunities to replicate this validation study in other settings. Even with a common statistical framework, the relative performance of these indirect approaches may be context specific, in terms of the bias-efficiency tradeoff as well as their ability to limit the influence of experimenter demand effects on survey responses. Ideally, applied researchers would have a body of validation studies to consult in choosing the method to use for their own research in a specific context.

References

Anderson, D. A., Simmons, A. M., Milnes, S. M., and Earleywine, M. (2007). Effect of response format on endorsement of eating disordered attitudes and behaviors. *International Journal of Eating Disorders* **40**, 1, 90–93.

Belli, R. F., Traugott, M. W., and Beckmann, M. N. (2001). What leads to voting overreports? contrasts of overreporters to validated votes and admitted nonvoters in the american national election studies. *Journal of Official Statistics* **17**, 4, 479–498.

Blair, G., Fair, C., Malhotra, N., and Shapiro, J. (2013). Poverty and support for militant politics: Evidence from Pakistan. *American Journal of Political Science* **57**, 1, 30–48.

Blair, G. and Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis* **20**, 1, 47–77.

Blair, G. and Imai, K. (2013). list: Statistical methods for the item count technique and list experiment. available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=list`.

Blair, G., Imai, K., and Lyall, J. (Forthcoming). Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science* .

Blair, G., Imai, K., and Zhou, Y.-Y. (2014a). Design and analysis of randomized response technique. Tech. rep., Princeton University.

Blair, G., Imai, K., and Zhou, Y.-Y. (2014b). rr: Statistical methods for the randomized response. available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=rr`.

Buchman, T. A. and Tracy, J. A. (1982). Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique. *Journal of Accounting Research* **20**, 1, 263–271.

Bullock, W., Imai, K., and Shapiro, J. N. (2011). Statistical analysis of endorsement experiments: Measuring support for militant groups in Pakistan. *Political Analysis* **19**, 4, 363–384.

Bursztyn, L., Callen, M., Ferman, B., Hasanain, S. A., and Yuchtman, N. (2014). A revealed preference approach to the elicitation of political attitudes: Experimental evidence on anti-americanism in pakistan. Working Paper 20153, NBER.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology* **39**, 5, 752–766.

Cialdini, R. (1993). *Influence: The Psychology of Persuasion.* Harper Collins, New York.

Corstange, D. (2009). Sensitive questions, truthful answers?: Modeling the list experiment with LISTIT. *Political Analysis* **17**, 1, 45–63.

Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research* **40**, 1, 169–193.

Dalton, D. R., Wimbush, J. C., and Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base-rates for sensitive behavior. *Personnel Psychology* **47**, 817–828.

Elffers, H., Robben, H. S., and Hessing, D. I. (1992). On measuring tax evasion. *Journal of Economic Psychology* **13**, 545–67.

Fair, C. C., Littman, R., Malhotra, N., and Shapiro, J. N. (2014). Relative poverty, perceived violence, and support for militant politics: Evidence from pakistan. Working Paper.

Folsom, R. E. (1974). A randomized response validation study: Comparison of direct and randomized reporting of dui arrests. *Chapel Hill, NC* Research Triangle Institute.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 4, 1360–1383.

Gingerich, D. W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis* **18**, 3, 349–380.

Glynn, A. N. (2013). What can we learn with statistical truth serum?: Design and analysis of the list experiment. *Public Opinion Quarterly* **77**, 159–172.

Gonzalez-Ocantos, E., de Jonge, C. K., Meléndez, C., Osorio, J., and Nickerson, D. W. (2012). Vote buying and social desirability bias: Experimental evidence from Nicaragua. *American Journal of Political Science* **56**, 1, 202–217.

Hessing, D. J., Elffers, H., and Weigel, R. H. (1988). Exploring the limits of self-reports and reasoned action: An investigation of the psychology of tax evasion behavior. *Journal of Personality and Social Psychology* **54**, 405–13.

Holbrook, A. L. and Krosnick, J. A. (2010a). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly* **74**, 2, 328–343.

Holbrook, A. L. and Krosnick, J. A. (2010b). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly* **74**, 1, 37–67.

Horvitz, D. G., Shah, B., and Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the American Statistical Association (Social Statistics Section)* 65–72.

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association* **106**, 494, 407–416.

Junger, M. (1989). Discrepancies between police and self-report data for dutch racial minorities. *British Journal of Criminology* **29**, 273–83.

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-semitism in germany: A comparison of randomized response and direct questioning. *Social Science Research* **41**, 6, 1387–1403.

Kuklinski, J. H., Cobb, M. D., and Gilens, M. (1997). Racial attitudes and the "New South". *Journal of Politics* **59**, 2, 323–349.

Kulka, R. A., Weeks, M. F., and Folsom, R. E. (1981). A comparison of the randomized response approach and direct questioning approach to asking sensitive survey questions. *Working paper, Research Triangle Institute, NC* .

LaBrie, J. W. and Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *Journal of Sex Research* **37**, 4, 321–326.

Lamb, C. W. and Stem, D. E. (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research* **15**, 616–21.

Lara, D., Strickler, J., Diaz Olavarrieta, C., and Ellertson, C. (2004). Measuring induced abortion in mexico: A comparison of four methodologies. *Sociological Methods and Research* **32**, 4, 529–558.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., and Maas, C. (2005). Meta-analysis of randomized response research. *Sociological Methods and Research* **33**, 319–348.

Locander, W., Sudman, S., and Bradburn, N. (1976). An investigation of interview method, threat, and response distortion. *Journal of the American Statistical Association* **71**, 269–275.

Lyall, J., Blair, G., and Imai, K. (2013). Explaining support for combatants during wartime: A survey experiment in Afghanistan. *American Political Science Review* **107**, 4, 679–705.

O'Keefe, D. J. (1990). *Persuasion: Theory and Research*. Sage Publications.

Petty, R. E. and Cacioppo, J. T. (1986). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag.

Petty, R. E., Cacioppo, J. T., and Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research* **10**, 135–146.

Petty, R. E. and Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. T. Fiske, and G. Lindzey, eds., *The Handbook of Social Psychology: Volume One*. John Wiley and Sons, 4th edn.

Public Policy Polling (2011). Toss up on mississippi 'personhood' amendment. Tech. rep., Public Policy Polling.

Shiraito, Y. and Imai, K. (2012). endorse: R package for analyzing endorsement experiments. available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=endorse`.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin* **133**, 5, 859–883.

Tracy, P. E. and Fox, J. A. (1981). The validity of randomized response for sensitive measurements. *American Sociological Review* **46**, 187–200.

Tsuchiya, T., Hirai, Y., and Ono, S. (2007). A study of the properties of the item count technique. *Public Opinon Quarterly* **71**, 2, 253–272.

van den Hout, A., van der Heijden, P. G., and Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis* **51**, 12, 6060–6069.

van der Heijden, P. G. M., van Gils, G., Bouts, J., and Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research* **28**, 505–537.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 309, 63–69.

Warner, S. L. (1971). The linear randomized response model. *Journal of the American Statistical Association* **66**, 336, 884–888.

Wolter, F. and Preisendorfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research* **42**, 3, 321–353.

Wood, W. and Kallgren, C. A. (1988). Communicator attributes and persuasion: Recipients' access to attitude-relevant information in memory. *Personality and Social Psychology Bulletin* **14**, 172–182.