What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment*

Adam N. Glynn†

July 23, 2010

Abstract

Due to the inherent sensitivity of many survey questions, a number of researchers have adopted an indirect questioning technique known as the list experiment (or the item count technique) in order to minimize bias due to dishonest or evasive responses. However, standard practice with the list experiment requires a large sample size, is not readily adaptable to regression or multivariate modeling, and provides only limited diagnostics. This paper addresses all three of these issues. First, the paper presents design principles for the standard list experiment (and the double list experiment) to minimize bias and reduce variance as well as providing sample size formulas for the planning of studies. Additionally, this paper investigates the properties of a number of estimators and introduces an easy-to-use piecewise estimator that reduces necessary sample sizes in many cases. Second, this paper proves that standard-procedure list experiment data can be used to estimate the probability that an individual holds the socially undesirable opinion/behavior. This allows multivariate modeling. Third, this paper demonstrates that some violations of the behavioral assumptions implicit in the technique can be diagnosed with the list experiment data. The techniques in this paper are illustrated with examples from American politics.

*An earlier version of this paper was presented at the 2010 Midwest Political Science Association conference. The author would especially like to thank Adam Berinsky for providing the list experiment data that was used in this paper as well as his collaboration in composing the items in the list experiment. These data were collected in conjunction with the 2010 Political Experiments Research Lab (PERL) Omnibus Study (Adam Berinsky PI). The author would also like to thank Matt Blackwell, Richard Nielsen, Justin Grimmer, Chase Harrison, Sunshine Hillygus, Gary King, and Kevin Quinn for their helpful comments and suggestions. The usual caveat applies.

†Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@fas.harvard.edu

1 Introduction

Along with sex, drugs, and religion, politics is an inherently sensitive subject. One can be embarrassed, ostracized, jailed, attacked, and even killed for one's political beliefs and actions. The sensitivity of politics has even been formally recognized in the widespread use of the secret ballot within modern democracies. Despite this acknowledged sensitivity, roughly half of the empirical political science articles published in leading journals rely on survey data (see footnote 1 from King et al. (2001)), even though these data (and the inferences made from them) are susceptible to inaccuracy when sensitive questions are included.

The inaccuracy due to sensitive questions is not trivial. For example, studies have shown that routinely a quarter or more of respondents who report voting actually did not vote (e.g., see Silver et al. (1986)). Furthermore, the effect of sensitive questions on our inferences may become more problematic once we consider the inclusion of "don't know" answer possibilities or the effects of item nonresponse (Berinsky, 1999).

One solution to this problem has been the use of indirect questioning techniques, which provide respondents some amount of privacy protection via randomization or aggregation. In the classic technique of randomized response (Warner, 1965), respondents flip a coin (or use some other randomization device) to decide which question to answer (without revealing to the researcher the result of the randomization or which question they answered). Due to known randomization probabilities, the researcher can estimate population proportions for the sensitive question, but does not know any individual's response to the sensitive question.

In aggregation techniques, respondents are typically asked how many of a list of questions apply to them. As long as the entire list does not apply, the respondent can be assured that the researcher does not know their answer to the sensitive question. Furthermore, if the lists are varied from respondent to respondent, the researcher can estimate population proportions for the sensitive question. One popular variant of aggregated response is known as block

total response (Raghavarao and Federer, 1979), of which a special case is known as the item count technique (Miller, 1984), the unmatched count technique (Dalton et al., 1994), or the list experiment (Sniderman and Carmines, 1997; Kuklinski et al., 1997a,b).

Recently, the list experiment has gained in popularity over randomized response due to a least two factors.[1] First, the list experiment only requires that respondents be able to answer how many items on a list apply to them, and therefore it is easier to conduct and easier to understand than randomized response, which requires respondents to flip a coin or to utilize some other randomization device.[2] This is particularly true for phone or internet surveys, where the respondent must provide their own randomization device in order for privacy protection to be assured. Second, recent experimental results have shown that the list experiment inspires more trust and acceptance among respondents and produces more reliable answers than randomized response techniques (Hubbard et al., 1989; Coutts and Jann, 2009).

However, while the list experiment provides an appealing alternative to direct questioning and randomized response, there are at least three major difficulties with the list experiment that have limited its use. First, the list experiment tends to require a large sample size and budget in order to achieve reasonable levels of precision (Tsuchiya, 2005; Tsuchiya et al., 2007; Corstange, 2009). For example, Streb et al. (2008) used 2,056 observations in their study to achieve only a 5% standard error. Second, standard analysis of the list experiment does not allow the analyst to diagnose many violations of the behavioral assumptions implicit in the technique. Third, it is difficult to use the standard-procedure list experiment in a regression or multivariate modeling framework.[3]

In this paper, I address these three difficulties by asking what we can learn from the

[1] For a summary of the uses of the list experiment within political science see Gonzalez-Octanos et al. (2010).

[2] Of course, even the task of answering how many items apply may be too difficult or annoying for many survey respondents (Tsuchiya et al., 2007).

[3] In ground breaking work, Corstange (2009) provides a method for multivariate modeling, but this requires an additional independence assumption and a procedural change in the administration of the list experiment.

standard-procedure list experiment without additional independence assumptions. I present design principles for the standard list experiment (and the double list experiment) to minimize bias and reduce variance as well as providing sample size formulas for the planning of studies. I also present some statistical properties of a number of estimators under certain behavioral assumptions. Finally, this paper provides techniques that allow the multivariate modeling of standard-procedure list-experiment data (i.e., experimental information on the sensitive item can be used as a left-hand side variable or a right-hand side variable in a regression) and the estimation of the probability that an individual holds the socially undesirable opinion.

2 The Standard List Experiment

The list experiment works by aggregating the sensitive item with a list of other non-sensitive items. For example, in a 2010 study,[4] 376 adult male respondents were randomized to a *baseline (or control)* group. These respondents were asked the following question:

```
   I am now going to give you a list of 4 statements.  Please tell me HOW MANY
of them are true for you.  I don't want to know which ones, just HOW MANY


  1. I have money invested in an Individual Retirement Account.
  2. I have sometimes been unable to pay my bills.
  3. I usually choose to buy organic foods.
  4. I usually shop at Wal-Mart.
```

In a separate, *treatment* group, 388 adult male respondents were randomized to receive the same list with the following sensitive item appended (although the list order was randomized):

[4]These data were collected in conjunction with the 2010 Political Experiments Research Lab (PERL) Omnibus Study (Adam Berinsky PI).

```
5. I would not vote for a woman for President.
```

For this example, a respondent in the treatment group who answers the "how many" question with anything less than "five" is provided a level of privacy because they could plausibly claim that they would not say yes to the sensitive item. Furthermore, if due to the question format, respondents in both groups answer the question honestly, then the analyst can estimate the true proportion in the population that would not vote for a female presidential candidate. This is often accomplished by taking the difference between the average response among the treatment group and the average response among the baseline group (i.e., a difference-in-means estimator). If the overall sample of respondents is randomly selected from the population of interest, and all of the other survey error components are negligible, then this estimator will be unbiased for the proportion in the population that would not vote for a woman president (a proof is provided in Appendix A). Furthermore, because estimation is accomplished by taking the difference in mean responses between two independent sets of respondents, the variance of the estimator can be calculated with the standard large-sample formula for a difference-in-means, and large sample confidence intervals can be formed in the usual fashion.

Of course, the accuracy of difference-in-means estimator depends on people's responses to both the baseline and treatment lists. Previous research has pointed to the importance of list experiment design for the bias and variance of the estimator.

3 Design for the List Experiment

3.1 Standard Design Advice

If respondents answer the list questions honestly, then the estimator will be unbiased, and therefore most list experiment designers have tried to create lists that will give respondents the privacy protection necessary to allow for honest responses. Most importantly, Kuklinski et al. (1997a) notes that "ceiling effects" can occur when a respondent would honestly respond

yes to all non-sensitive items. When this happens to a treatment group respondent, they no longer have the protection to honestly report their response to the sensitive item, and therefore this respondent may underreport their response to the treatment list. Kuklinski et al. (1997a) also notes a stark example of this problem. Their results for Non-Southern respondents showed that a large portion of their control group reported all of the non-sensitive items, and due to the consequent ceiling effects, the results in the the paper implied (non-sensically) a negative proportion for the sensitive item. Furthermore, near-ceiling effects may also be possible in addition to ceiling effects. For example, respondents who would report three non-sensitive items (with a four item baseline list) may underreport the number of items on the treatment list because they do not want to show even a likelihood of holding the socially undesirable opinion or behavior.

The concerns over ceiling effects and a lack of privacy protection have led to three generally accepted pieces of design advice. First, high prevalence non-sensitive items, which would increase the occurrence of ceiling effects, should be avoided (Droitcour et al., 1991). Second, low prevalence non-sensitive items should be avoided (at least in abundance). If respondents are aware that all the non-sensitive items have low prevalence, they may become concerned about the level of privacy protection and underreport their answers (Tsuchiya et al., 2007).[5] Third, lists should not be too short because short lists will also tend to increase the likelihood of ceiling effects (Kuklinski et al., 1997a).

Unfortunately, these three pieces of design advice tend to lead to increased variability in the responses to the "how many" questions and therefore increased variance of the estimator (Tsuchiya et al., 2007; Corstange, 2009).[6] This state of affairs presents the list designer

[5]In addition to concerns about perceived privacy protection, a number of authors have commented on the need for the individual baseline items to be credible in comparison to the sensitive item. Kuklinski et al. (1997a) warns against contrast effects (where the resonance of the sensitive item overwhelms the baseline items), and Droitcour et al. (1991) advises that the non-sensitive items should be on the same subject as the sensitive item. In short, the standard advice is to choose the baseline list such that the respondent is unlikely to report all non-sensitive items, and to choose baseline items that do not seem out of place with the sensitive item.

[6]In addition to increased variance, Tsuchiya et al. (2007) presents evidence that longer lists may produce

with an unfortunate tradeoff between ceiling effects and variable results (an apparent bias-variance tradeoff). However, while this bias-variance tradeoff will be unavoidable without at least a probabilistic reduction in privacy protection, we may be able to simultaneously minimize ceiling effects and response variability without reducing the perception of probabilistic privacy protection.

3.2 Reducing Bias and Variance with Negative Correlation

As summarized in the previous section, when designing a list experiment there are at least two goals we would like to achieve. First, we would like to limit ceiling effects and the bias that results from these effects. Second, we want to minimize the variance of the estimator. However, even if we accept that the list must not be too short, and that the list should not be largely composed of high or low prevalence items, there are still a number of design options available to increase the accuracy of the estimator.

First, the designer can allocate different sample sizes to the treatment and baseline groups. However, the benefits from this design option appear to be minimal in most cases. For example, the list experiment described in Section 2 produced an estimated variance for the baseline list of 0.71 and an estimated variance for the treatment list of 1.04. If we assume that these estimates are accurate, and we hold the overall sample size fixed at 764, then it would have been optimal to use 346 of these observations for the control group instead of the 376 that were used. However, the standard error reduction that would have been achieved by changing from equal sample sizes to the optimal allocation is minimal ($\approx .0003$). Furthermore, the optimal allocation cannot be obtained prior to the experiment without the invocation of strong untestable assumptions (see Appendix B for details). Finally, as shown in the next section, there are some benefits to using equal sample sizes in each group when the double list estimator is utilized.

measurement error due to the inability of respondents to remember their responses to all items on the list.

While the optimal allocation of sample sizes typically provides limited benefits, the choice of baseline items provides the potential for more substantial variance reduction. If items are chosen for the baseline list so that negative correlation between the responses to these items limits the variability of the baseline "how many" responses, then the list can be kept relatively short while minimizing the likelihood of ceiling effects (and hence bias due to these effects). For example, consider the baseline list presented in the previous section. Notice that the two pairs of items largely accomplish our design objectives. For the first and second items, investing in an IRA is more likely to be true for high income respondents, while being unable to pay bills is more likely to be true for low income respondents, and both items will be true for only a small number of respondents. Similarly, the third and fourth items are paired to be negatively correlated. Respondents who usually choose to buy organic food are not likely to shop at Wal-Mart often. Therefore, the individual items have a non-trivial prevalence, but we expect few individuals to report all four baseline items, and bias due to ceiling effects should be minimized. Furthermore, while the reduction in variation for the baseline list represents a probabilistic reduction in privacy protection for the respondents, it is unlikely that most respondents will perceive the negative correlation designed into the list, and therefore this list should be less likely to induce underreporting than would a list containing four low prevalence items.

The benefits of negative correlation design can be seen in the list example presented in Section 2. Table 1 presents the results from this list experiment, where the difference-in-means estimate of the percent of men who would not vote for a woman for president is 19%, and the standard error is 6.8%. Notice that this standard error is quite good for a list experiment of this sample size. For example, consider the list experiment with four baseline items utilized by Streb et al. (2008) to assess American support for a female president. This list was adapted by Kane et al. (2004) from the canonical three item list developed for the 1991 National Race and Politics Survey (Sniderman et al., 1992) and used in Kuklinski

Table 1: Results from the list experiment described in Section 2.

	Baseline condition	Treatment condition	Diff-in-means Estimate
	1.59	1.78	0.19
	(0.043)	(0.052)	(0.068)
n =	376	388	

et al. (1997a) and Kuklinski et al. (1997b). The Streb et al. (2008) piece, using the adapted canonical list and attempting to assess support for a female president, produced a standard error of 7.2% with a sample size of 513 males in the baseline group and 512 males in treatment group. To put this in perspective, the list experiment in Section 2 produced a smaller standard error than the canonical four-item list for a similar sensitive question with only 3/4 of the sample size. Although the comparison is not exact, this establishes the promise of the technique. The genesis of this list experiment is discussed in more detail in Berinsky and Glynn (2010).

While the negative correlation between pairs of baseline items can minimize the bias for the treatment list and the variance of responses for the baseline list, it will not guarantee low variance for the difference-in-means estimator because the covariance between the baseline list and the sensitive item may be large. Additionally, while it is straightforward to calculate the sample size necessary to guarantee a maximum half width for a large sample confidence interval (see Appendix A for details), these sample sizes can be quite conservative because this calculation depends on both the unknown prevalence of the sensitive item and the unknown correlation between the sensitive item and the responses to the baseline list. Finally, the single list experiment may be not be ideal because only respondents in the treatment group receive the sensitive item, and therefore the sample size is effectively cut in half. The double list experiment, described in the next section is one way to alleviate both of these deficiencies.

3.3 Design for the Double List Experiment

Droitcour et al. (1991) shows that we can present the sensitive item to all respondents by using two baseline lists. For example, denote the baseline list presented in Section 2 as baseline list A, and denote the following list as baseline list B.

```
Baseline List B:


1. I have money invested in a mutual fund or individual stocks.
2. I rent my home.
3. I exercise at least four times a week.
4. I own a gun.
```

In the double list experiment, the respondents are again separated randomly into two groups, but both groups function simultaneously as baseline and treatment groups. For example, in the previously cited study, 376 adult male respondents first received the A baseline list and then received the B baseline list with the sensitive item appended. The 388 adult male respondents in the other group received the B baseline list and then received the A baseline list with the sensitive item appended. In this way the first group functions as the treatment group with respect to the B baseline list, while the second group functions as the treatment group with respect to the A baseline list, and these two list experiments both provide difference-in-means estimators that can be averaged in order to maximize precision (see Appendix C for details).

The negative correlation design from the previous section can be applied to both baseline lists in the double list experiment, however, there are additional design opportunities presented by the double list experiment. Specifically, we can increase our certainty about the sensitive item by using two baseline lists that have positive correlation on the responses to the "how many" quesiton (see Appendix C for details). To understand why this works,

consider the extreme case where the two baseline lists are identical. In this case, we would effectively be directly asking each respondent the sensitive question. Positive correlation between the lists is an attempt to approach this level of precision.

As an example, consider that the first item from the A baseline list, "I have money invested in an Individual Retirement Account," is similar to the first item on the B baseline list, "I have money invested in a mutual fund or individual stocks," and responses to these items are likely to be positively correlated. This relationship holds for all four items on the A and B lists, so we can expect that positive correlation will be induced between the responses to the two baseline lists. In fact, due to positive correlation between the lists, the double list experiment described above produced a standard error of 6.3% with the 764 respondents. This represents an improvement of half a percent in the standard error, although the improvement might be greater with a different baseline B list. The B list presented here underperformed compared to the baseline A list, producing a single list standard error of 7.2% and a baseline list variance of 0.95.

Perhaps more importantly, the double list experiment allows for more precise planning of future studies if the baseline items for both lists are pre-tested. If equal sample sizes are used for each group, and respondents react in the same way to the inclusion of the sensitive item on the two lists, then the variance of the average of the two difference-in-means estimators will depend only on the variance of the two baseline lists, the covariance between the two baseline lists, and the variance of the sensitive item. Therefore optimal design of the baseline lists does not depend highly on the correlation between the sensitive item and the baseline list, and the variance of the estimator will be minimized when negative correlation within each baseline list minimizes the variance of each baseline list and positive correlation between each list maximizes the covariance between the lists (see Appendix C for a proof). Pre-testing can establish plausible values for most of the quantities in the variance, and the variance of the sensitive item is bounded above by .25, so we can obtain sample size

Table 2: Data from the list experiment described in Section 2.

Estimated Proportions	Source	Number of Reported Items 0	1	2	3	4	5	Sum
Row 1	Treatment List	0.088	0.317	0.387	0.157	0.041	0.010	1.00
Row 2	Proportion at Least	1.000	0.912	0.595	0.208	0.051	0.010	-
Row 3	Baseline List	0.085	0.372	0.420	0.109	0.013	0.000	1.00
Row 4	Proportion at Least	1.000	0.914	0.542	0.122	0.013	0.000	-
Row 5	Row 2 - Row 4	0.000	-0.002	0.053	0.086	0.038	0.010	0.186

calculations that are not as conservative as the sample size calculations for the single list experiment (see Appendix C for the formula).

This section and the previous section have demonstrated that smart design can nearly eliminate ceiling effects and can dramatically reduce the standard error of estimates for population proportions. Therefore, these techniques increase the probability of achieving significant results (i.e., increase power), and decrease necessary sample sizes (i.e., reduce costs). However, we can extract further information from the list experiment by more closely analyzing the resulting data.

4 Analysis of the List Experiment

4.1 The List Experiment Data and the Joint Proportion Estimates

While the difference-in-means analysis provides an unbiased estimator for the population proportion of the sensitive item, there is potentially useful information available in the list experiment data that is ignored by the difference-in-means approach. This can be most easily seen by examining the table of responses to the treatment and baseline lists.

Table 2 reports the data from the Section 2 list experiment as estimated proportions. Notice that we can reconstruct the difference-in-means estimator in the following manner. For each list and for each possible number of indicated items, estimate the proportion reporting at least that number (rows 2 and 4 of Table 2). For each possible number of indicated items,

subtract the proportion "at least" on the baseline list from the proportion "at least" on the treatment list. Finally, notice that if we sum these differences, we obtain the difference-in-means estimator.

This result is not surprising. The procedure described above is just an arithmetic alternative for obtaining a difference-in-means. However, the table also provides information that is not being incorporated by the difference-in-means estimator. To see this, note that due to the treatment randomization, the proportion of treatment group respondents reporting five items on the treatment list (0.010) estimates the proportion in the population that would say yes to the sensitive item (if forced to be honest) and yes to all four non-sensitive items.

Similarly, the proportion of treatment group respondents reporting at least four items on the treatment list ($0.041+0.010 = 0.051$) minus the proportion of baseline group respondents reporting all four non-sensitive items (0.013) estimates the population proportion that would honestly say yes to the sensitive item and yes to exactly three non-sensitive items (0.038). To see this, note that the proportion of treatment group respondents reporting at least four items on the treatment list (0.051) estimates the population proportion that would honestly say yes to the sensitive item and yes to exactly three non-sensitive items *or* that would honestly say no to the sensitive item and yes to all four non-sensitive items (0.041), plus the population proportion that that would honestly say yes to the sensitive item and to all four non-sensitive items (0.010). Analogously, the proportion of baseline group respondents reporting at least four items on the baseline list estimates the population proportion that would honestly say yes to the sensitive item and to all four non-sensitive items *or* that would honestly say no to the sensitive item and yes to all four non-sensitive items (0.013). Therefore, if we take the difference between these two proportions ($0.051 - 0.013 = 0.38$), we are left with an estimate of the population proportion that would honestly say yes to the sensitive item and yes to exactly three non-sensitive items.

The other entries in the fifth row of Table 2 can be similarly interpreted as estimates

of joint population proportions that would honestly say yes to the sensitive item and yes to exactly k non-sensitive items, where k is equal to one less than the "number of reported items" for each entry (Appendix D provides a proof that the list experiment data identifies these joint proportions).[7] The remainder of this paper explores the potential uses of these joint proportions and the other information in this piecewise table.

4.2 Conditional Analysis

The estimated joint proportions in Table 2 provide the basis for a number of different conditional analyses. The most obvious of these is an analysis estimating the probability that an individual holds the sensitive belief/trait, conditional on their answer to either the treatment or baseline list. For example, the 0.038 in the fifth row of Table 2 indicates that 3.8% of the respondents in the population hold the sensitive trait and exactly three of the non-sensitive traits. Therefore, for the 10.9% of the baseline group that reports three items, we can estimate the probability of their holding the sensitive trait as 3.8%/10.9% = 34.9%. Similarly, for the 4.1% of the individuals in the treatment group that indicate four items in their response, we can estimate the probability of their holding the sensitive trait as 3.8%/4.1% = 92.7%. These conditional probabilities, reported in Table 3, are the fundamental building blocks for treating the sensitive item as missing data (e.g. multiple imputation, the EM algorithm, or data augmentation), which allows multivariate modeling of the sensitive item.[8] These probabilities also provide an estimate of the amount of privacy protection (or lack thereof) for each possible response to the treatment and baseline lists. The formal definition of these conditional probabilities and this estimator is presented in

[7] For this table, the estimate for the proportion that indicate the sensitive item and none of the non-sensitive items is -0.002. As with the difference-in-means estimator (which can also produce negative estimates), this result either indicates a lack of comparability between the treatment and control groups (due to an insufficient sample size), or it indicates that at least some respondents are misrepresenting their answers on the "how many" questions.

[8] Clearly the negative estimated probabilities are nonsensical and are again due to a small sample size or misrepresentation.

Table 3: Estimated conditional probabilities from the list experiment described in Section 2.

Estimated	Number of Reported Items					
Probabilities	0	1	2	3	4	5
On Baseline List	-0.024	0.142	0.205	0.349	0.769	-
On Treatment List	0.000	-0.006	0.137	0.548	0.927	1.000

Appendix E.

There is also a simple piecewise regression estimator that can be created using the joint proportions.[9] To see this, note that if we create dummy variables to correspond with the "at least" rows of Table 2 (rows 2 and 4), then by regressing these dummy variables on the explanatory variables, we can piece together the desired regression curve by taking the differences between the row 2 and row 4 curves and and then summing. This algorithm is presented in Appendix E.

For example, suppose we want to investigate the relationship between racial resentment among males for a female presidential candidate. If we create a dummy variable to indicate when a treatment group respondent has indicated all five items, and we regress this variable (using logistic regression) on the racial resentment variable, then we can estimate the logistic regression curve that describes the predicted probability that a white male (at different levels of racial resentment) would not vote for a woman *and* would indicate all four non-sensitive items. Similarly, we can create a dummy variables to indicate when treatment and baseline group respondents have indicated at least four items. If we separately regress each of these dummy variables on the racial resentment variable, the difference between the curves describes the predicted probability that a white male (at different levels of racial resentment) would not vote for a woman *and* would indicate three non-sensitive items. This process can be repeated with dummy variables for at least three items, at least two items, and at least

[9]Of course, if the right hand side variables are categorical, it is also possible to simply perform the difference-in-means analysis within the different strata. Additionally, Tsuchiya (2005) provides an alternative conditional estimator for categorical conditioning variables that will provide greater precision when an additional independence assumption holds.

one item. The resulting curves can then be added to produce the predicted probability that a male (at different levels of racial resentment) would not vote for a woman.

This curve (with 95% pointwise bootstrapped confidence intervals) is presented in Figure 1. Notice that as we would expect, the probability of a white male not voting for a woman increases as racial resentment increases and this probability is significantly different from zero at larger values of racial resentment. At the low end of racial resentment, the curve and the confidence interval were truncated at zero in order to avoid nonsensical estimates, but this truncation did not affect the estimates or the confidence intervals for larger values of racial resentment.

4.3 Beyond the Difference-in-Means Estimator

While the previous section demonstrated that the information in Table 2 can be used to perform conditional analysis, the extra information in the table (typified by the negative number in the fifth row) suggests alternative estimators for the overall population proportion of the sensitive item. These estimators may be preferable to the difference-in-means estimator under some circumstances, however, the choice of estimators in this framework depends on a number of factors.

When discussing alternatives to the difference-in-means estimator, the first thing to note is that the difference-in-means estimator, while being unbiased, is inadmissible. It can produce negative estimates or estimates greater than one, and therefore a truncated difference-in-means estimator, which forces estimates to be in the zero-one range, will always be closer to the true proportion.[10] However, while the truncated difference-in-means estimator is clearly preferable to the difference-in-means estimator, it is important to note that truncation may produce bias when the true parameter is near the zero boundary (as we sometimes expect with socially undesirable items).

[10]In practice, researchers tend to use the truncated estimator (see for example the zero reported in Table 1 of Kuklinski et al. (1997b)).

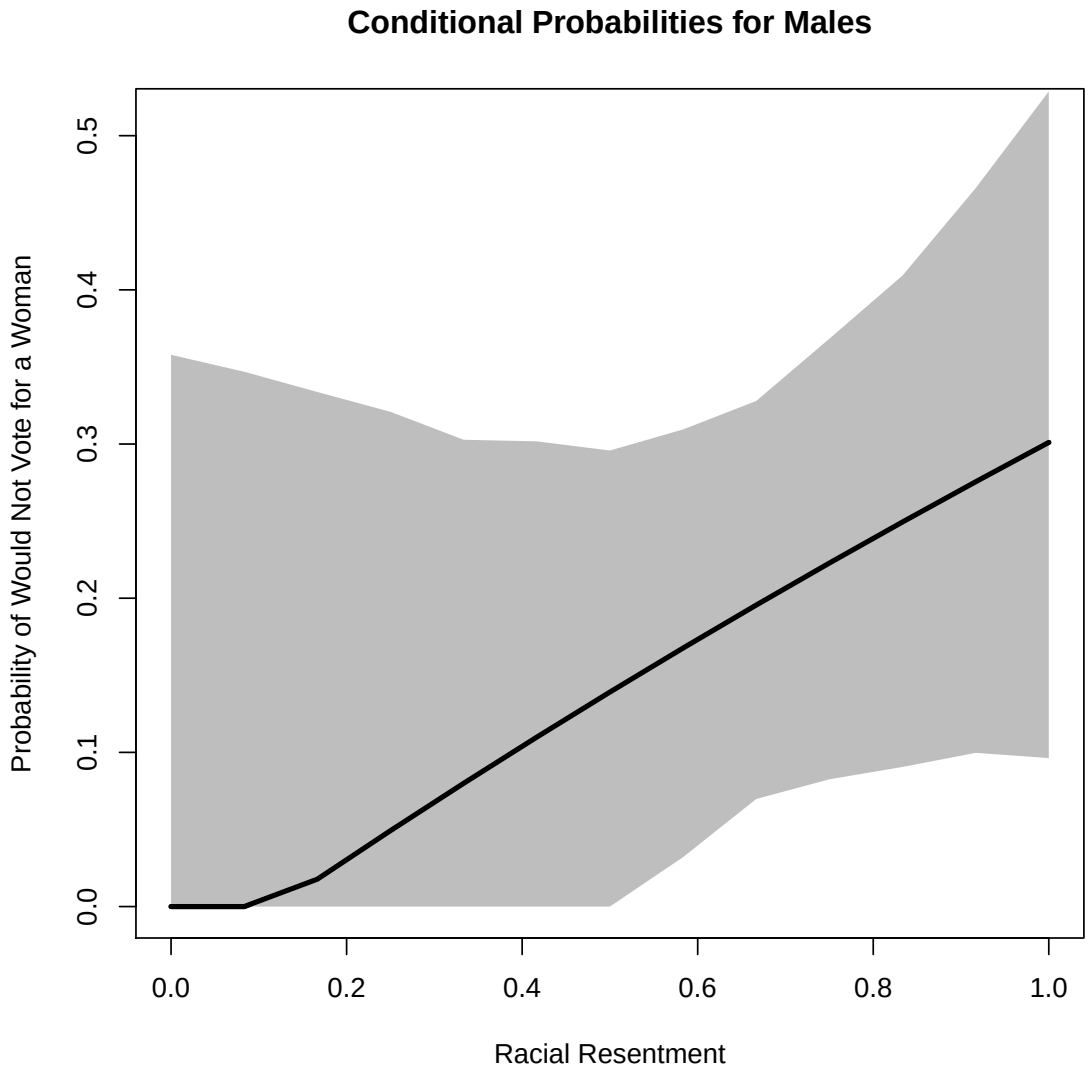

Figure 1: Piecewise regression estimate for the probability of not voting for a woman, conditional on the racial resentment measure. The shaded area represents 95% pointwise bootstrapped confidence intervals.

Because bias is somewhat unavoidable, choosing an estimator will involve a tradeoff between small bias (often induced by some kind of truncation) and variance (often induced by not efficiently using all the information available in Table 2). In addition to the truncated difference-in-means estimator, we might try a piecewise estimator that adjusts the joint proportion estimates in Table 2 before summing them to estimate the overall proportion.

A simple piecewise estimator would make the following two adjustments to the entries in the fifth row of Table 2 (before these values were summed to produce an estimate). First, any negative values would be changed to zero. Second, any values greater than the third row values (one column to the left) would be reduced to the corresponding third row value. This second adjustment ensures that some conditional probabilities will not exceed one. For the list experiment data in Table 2, these adjustments do not change the estimate much (only the -0.002 in the fifth row is changed to zero) and the overall estimate only increases to 0.188. However, in some cases, this adjustment can be dramatic (Glynn et al., 2010).

Another estimator to consider is the maximum likelihood estimator (Imai (2010) develops this along with a regression estimator). Maximum likelihood estimators are known to have a number of good properties, and may provide better estimates of the population proportion for this problem. The piecewise estimator suggests a different parameterization of the log-likelihood, and this is presented in Appendix F.

The simulation study pictured in Figure 2 presents some evidence as to the properties of these estimators (based on 1,000 Monte Carlo simulations). The data for the baseline list responses were generated using a multinomial distribution with parameters set according to the estimated baseline proportions in Glynn et al. (2010).[11] The sensitive item was independently generated as a bernoulli random variable with probability π_K.[12] The trun-

[11] Of the baseline respondents in this study, 28% report zero items, 54% report one item, and 18% report two items.

[12] Simulations where the probability of the bernoulli random variable depended on the implied baseline response did not substantively change the findings.

cated difference-in-means estimates (—) and the piecewise estimates (− − −) were calculated using the procedures described above. The maximum likelihood estimates (. . .) were calculated with an EM algorithm as suggested in Imai (2010) and adapted for the multinomial distribution.

Two impressions are immediately apparent from the bias and root mean square error (RMSE) plots in Figure 2. First, the truncated difference-in-means estimator has less bias than the piecewise estimator, which has less bias than the MLE. This relationship holds for all sample sizes and parameter values considered. Second, the comparison in terms of RMSE depends on the parameter value. When the sensitive item has a very small prevalence (0.01), the truncated difference-in-means estimator is preferred to the piecewise estimator, which is preferred to the MLE. When the sensitive item has a larger prevalence (0.05 or 0.10), the MLE is preferred in comparison to the difference-in-means estimator and is very similar to the piecewise estimator.

The differences between the difference-in-means estimator and the piecewise estimator are expected given the analysis in Table 2. The difference-in-means estimator is the sum of the entries in the fifth row of that table (a negative sum would be truncated at zero). For the small parameter values in this simulation, the piecewise estimator tends to differ from the difference-in-means estimator when some of the entries in the fifth row are negative (e.g., the $-.002$ value discussed above). This means that when the piecewise estimator differs from the difference-in-means estimator, it will be larger, and will therefore have more positive bias (when respondents are assumed to be honestly reporting their answers). However, by constraining the estimated joint probabilities in Table 2 to be logically consistent, the piecewise estimator has less variance than the difference-in-means estimator, and therefore will have smaller RMSE in cases where the bias is not too large (e.g. $\pi_K \geq .05$).

The relationship between the piecewise estimator and the MLE is more complicated. Intuitively, the piecewise estimator is a two step estimator that first uses the data from

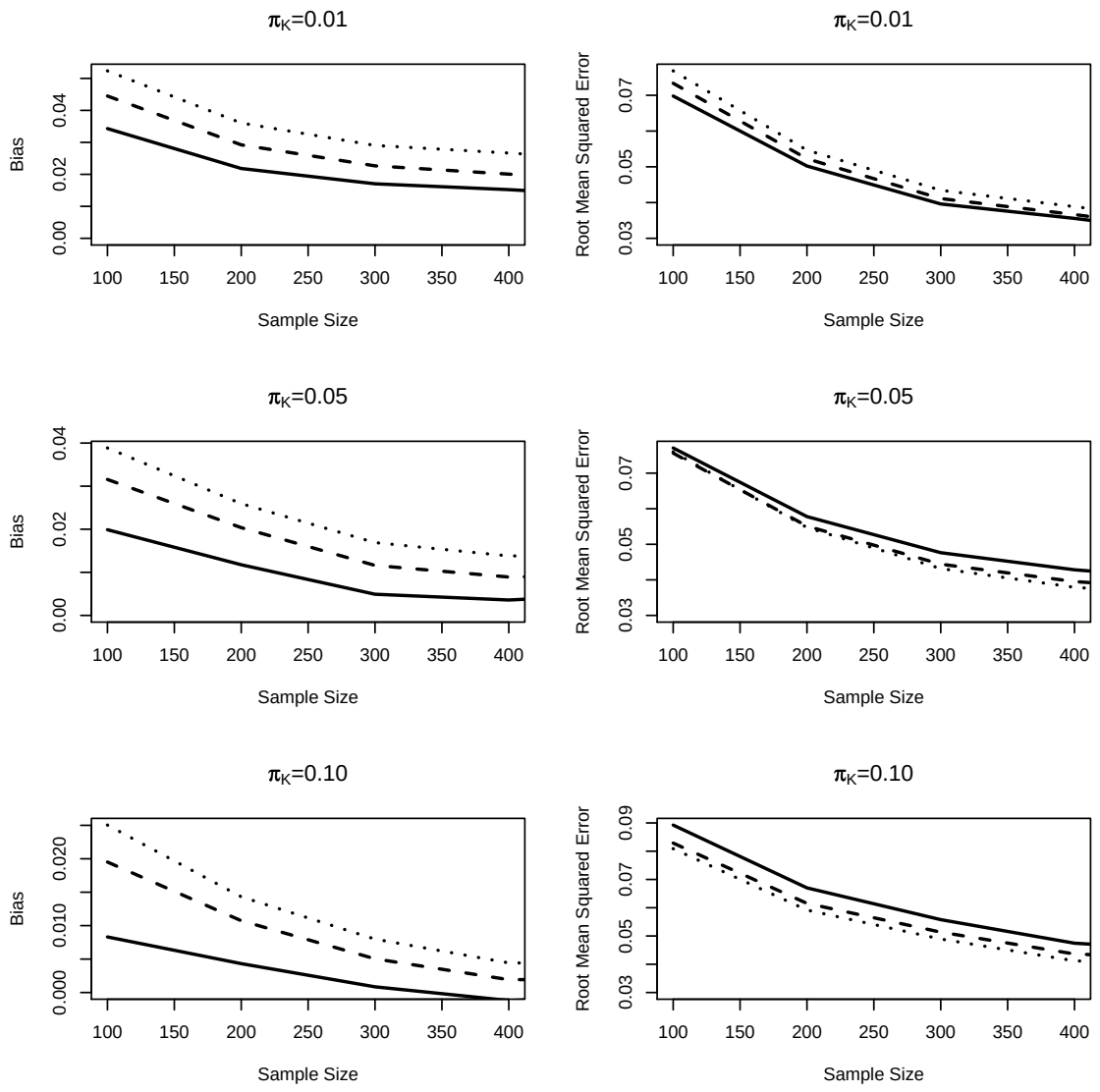

Figure 2: Comparison of bias and root mean squared error for the truncated difference-in-means estimator (—), the piecewise estimator ($- - -$), and the maximum likelihood estimator (. . .) at different levels of the true population proportion of the sensitive item (π_K) and different sample sizes. For each parameter value and sample size, 1,000 Monte Carlo simulations were used to generate the data. The baseline list responses were generated using a multinomial distribution with parameters set according to the estimated baseline proportions in Glynn et al. (2010) (28% reported zero items, 54% reported one item, and 18% reported two items). The sensitive item was independently generated as a bernoulli random variable with probability π_K.

the baseline group to estimate the parameters for the baseline list, and then backs out the sensitive item parameter using the data from the treatment group. The MLE uses the data from both the baseline and the treatment groups to simultaneously estimate all parameters. However, it is useful to note that the piecewise estimator, while extremely easy to calculate, performs nearly as well as the MLE in terms of RMSE and has smaller bias for these simulations.

4.4 Detecting Violations of the Behavioral Assumptions

Until now, we have assumed that respondents answer the list questions honestly. However, this is a strong assumption because almost no list experiment completely eliminates ceiling effects. Even for the list experiment presented in Table 2, 0.013 of the baseline group respondents reported all four of the non-sensitive items. We might worry that some respondents in the treatment group would have said yes to all four of the non-sensitive items and were therefore directly asked the sensitive item. Some of these respondents might have underreported the answer to the treatment list question.

Furthermore, respondents near the ceiling (e.g., would say yes to three non-sensitive items) might be worried about appearing likely to hold the socially undesirable trait/opinion. These respondents might also underreport their responses to the baseline list. Finally, some respondents will be so hesitant about appearing to possibly hold the undesirable trait/opinion, that they may report zero on the treatment list regardless of what they would have reported on the baseline list.

It is possible, but unlikely, that one could detect such underreporting with a difference-in-means analysis. A negative difference-in-means may indicate underreporting, but it could also reflect imbalance between the treatment and baseline groups due to the sample size. Furthermore, even with an infinite sample size, if every respondent who would say yes to the sensitive item were to underreport by one, on average the difference-in-means would be zero,

and we would be unable to tell the difference between underreporting and zero prevalence of the sensitive item.

Fortunately, the fifth row elements of the piecewise table provide a more powerful test for underreporting. If we assume that baseline group respondents answer the question honestly, and we assume that treatment group respondents only misreport their answers by underreporting, then any negative value in the fifth row of the table either indicates underreporting or an imbalance between the treatment and baseline groups. Therefore, a hypothesis test may be used to find a significantly negative number in the fifth row of the piecewise table. If the null hypothesis of a zero can be rejected in favor of a negative proportion, then we can claim to have detected underreporting.

To avoid issues of multiple testing, a simple and reasonably powerful test for underreporting would test the proportion for those that would would say yes to the sensitive item and yes to all but one of the non-sensitive items. For the example in Table 2 this would involve testing whether 0.038 was significantly smaller than zero (obviously this test fails to reject). However, in general we might expect this test to have reasonably good properties. First, we expect underreporting to occur due to ceiling effects or near ceiling effects, which would imply that most likely the proportions reporting five (and maybe four) on the treatment list will be underreported. Second, we can't know the magnitude of each underreport. For example, an underreporter who would honestly report four on the treatment list could report three but might report zero instead. The proposed test would catch either of these underreports, while tests based solely on other entries in the fifth row of the table would not.

As an example, Glynn et al. (2010) finds a difference-in-means estimate of 0.061 but the estimate for the proportion that would say yes to the sensitive item and yes to all but one of the non-sensitive items is -0.046 with a standard error of 0.022 (the elements in the fifth row of the piecewise table are differences in proportions so the calculation of standard errors is straightforward). Therefore, if we believe that the treatment group never overreports and

the baseline group never misreports, then this is significant evidence of underreporting.[13]

Finally, it is useful to note that the proposed diagnostic tests utilize the fifth row of the piecewise table, and due to the fact that the piecewise estimator adjusts negative values in the fifth row to zero, this estimator will be more robust to ceiling effects than the truncated difference-in-means estimator.

5 Conclusion

The list experiment represents a powerful and increasingly popular tool for indirectly asking sensitive questions on surveys. However, there are at least three major difficulties with the list experiment that limit its efficacy. First, the list experiment tends to require a large sample size. Second, standard analysis does not allow the analyst to feasibly diagnose violations of the behavioral assumptions implicit in the technique. Third, it is difficult to use the standard-procedure list experiment in a regression or multivariate modeling framework.

This paper addresses all three of these issues. First, the design principles of negative within list correlation and positive between list correlation were introduced and sample size formulas were derived. These tools should improve the planning of future studies and increase the feasibility of the technique by reducing necessary sample sizes and increasing the precision of estimates. Additionally, the easy-to-use piecewise estimator was introduced and this estimator (as well as the maximum likelihood estimator) were shown to have greater precision than the traditional difference-in-means estimator at the cost of greater bias. When the true prevalence of the sensitive item is not too close to zero, these estimators appear to be preferable to the difference-in-means estimator (at least in terms of squared error loss).

Second, this paper developed a simple test for diagnosing violations of the assumption that respondents honestly report their answers to the treatment list questions. The structure

[13]Contrary to these assumptions, Glynn et al. (2010) finds evidence of overreporting by using a placebo test.

of this test also shows that the piecewise estimator will be more robust to ceiling effects than the difference-in-means estimator.

Third, this paper demonstrates that conditional analysis can be performed with the standard procedure list experiment by proving that the joint proportions and the conditional proportions can be estimated with list experiment data. These proportions allow regression analysis, the treatment of the sensitive item as missing data, and a measure for the implicit privacy protection provided by the list experiment.

However, while this paper has demonstrated the potential benefits from taking design and analysis seriously within the standard framework of the list experiment, there are undoubtedly other related aggregated response designs that might improve the reliability of our information. For example, if the sensitive question is asked directly prior to the administration of the list experiment, then respondents who answer with the socially undesirable response will provide a lower bound for the prevalence. Therefore, we need only utilize the list experiment (and all advice from the previous sections) on those individuals that respond with the socially desirable opinion when directly questioned on the sensitive item.

Furthermore, while this paper has started to address violations of the behavioral assumptions implicit in the list experiment. Many of these assumptions remain untested in much of the literature. Future research should focus on developing procedures for testing these assumptions.

Appendix A: Bias and Variance of the Difference-in-Means List Experiment Estimator

To formalize notation, the simple list experiment uses two lists: a baseline list with $K-1$ items, and a treatment list that includes the $K-1$ baseline items plus the sensitive item. We will assume that the individuals in the experiment are representative of the population, independent from each other, and that individuals are randomly assigned to the baseline

and treatment groups. Furthermore, we define the following random variables for items $k = 1, \ldots, K$, and individuals $i = 1, \ldots, n_1$ for the baseline group, $i = n_1 + 1, \ldots, n$ for the treatment group, where $n_2 = n - n_1$ is the number of individuals in the treatment group:

$$\tilde{y}_{ik} = \begin{cases} 1 & \text{if individual } i \text{ would say yes to item } k \text{ (if forced to be honest)} \\ 0 & \text{if individual } i \text{ would say no to item } k \text{ (if forced to be honest)} \end{cases}$$

These unobservable opinions are distributed as Bernoulli random variables (dependent within individuals), where π_k is the proportion of individuals in the population who would honestly say yes to item k. Therefore, the population proportion for the sensitive item (π_K) is the parameter of interest.

We further define sums over these random variables and parameters,

$$\begin{aligned} \tilde{y}_{i+}^{K} &= \sum_{k=1}^{K} \tilde{y}_{ik} \qquad & \tilde{y}_{i+}^{K-1} &= \sum_{k=1}^{K-1} \tilde{y}_{ik} \\ \pi_{+}^{K} &= \sum_{k=1}^{K} \pi_k \qquad & \pi_{+}^{K-1} &= \sum_{k=1}^{K-1} \pi_k \end{aligned}$$

where $\tilde{y}_{i+}^{K}$ is the total number of items that individual i would honestly say yes to that includes the sensitive item K, $\tilde{y}_{i+}^{K-1}$ is the total number of items that individual i would honestly say yes to, not including the sensitive item K, and π_{+}^{K} and π_{+}^{K-1} are their respective expected values.

In contrast to these unobservable opinions, we observe responses to the "how many" questions, y_{i+}^{K-1} for the $i = 1, ..., n_1$ individuals in the baseline group, and y_{i+}^{K} for the $i = n_1 + 1, \ldots, n$ individuals in the treatment group. The difference-in-means list experiment estimator can therefore be written as the following:

$$\hat{\pi}_K = \frac{1}{n_2} \sum_{i=n_1+1}^{n} y_{i+}^{K} - \frac{1}{n_1} \sum_{i=1}^{n_1} y_{i+}^{K-1} \tag{1}$$

Assumption 1 (Honest Responses) *$y_{i+}^{K-1} = \tilde{y}_{i+}^{K-1}$ for $i = 1, ..., n_1$, and $y_{i+}^{K} = \tilde{y}_{i+}^{K}$ for $i = n_1 + 1, \ldots, n$.*

If we utilize Assumption 1 and assume that all respondents to the list experiment honestly report their answers, then (1) will be an unbiased estimator of π_K:

$$\begin{aligned} E[\hat{\pi}_K] &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} E[y_{i+}^K] - \frac{1}{n_1}\sum_{i=1}^{n_1} E[y_{i+}^{K-1}] \\ &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} E[\tilde{y}_{i+}^K] - \frac{1}{n_1}\sum_{i=1}^{n_1} E[\tilde{y}_{i+}^{K-1}] \\ &= \frac{n_2}{n_2}\pi_K + \left(\frac{n_2}{n_2}\pi_+^{K-1} - \frac{n_1}{n_1}\pi_+^{K-1}\right) \\ &= \pi_K. \end{aligned}$$

Furthermore, the variance of (1) can be derived in terms of the variances of the sums,

$$\begin{aligned} V[\hat{\pi}_k] &= \frac{1}{n_2^2}\sum_{i=n_1+1}^{n} V[y_{i+}^K] + \frac{1}{n_1^2}\sum_{i=1}^{n_1} V[y_{i+}^{K-1}] \\ &= \frac{1}{n_2^2}\sum_{i=n_1+1}^{n} V[y_{i+}^K] + \frac{1}{n_1^2}\sum_{i=1}^{n_1} V[y_{i+}^{K-1}] \\ &= \frac{1}{n_2}V[y_{i+}^K] + \frac{1}{n_1}V[y_{i+}^{K-1}] \qquad (2) \\ &= \frac{1}{n_2}\left(V[y_{i+}^{K-1}] + V[y_{iK}] + 2\cdot Cov[y_{i+}^{K-1}, y_{iK}]\right) + \frac{1}{n_1}V[y_{i+}^{K-1}], \qquad (3) \end{aligned}$$

where (2) is the standard difference-in-means variance formula, and the decomposition in (3) can be used for design considerations. In (3), $y_{iK} \equiv y_{i+}^K - y_{i+}^{K-1}$ is defined to be the difference between what a respondent would have reported on treatment list and what they would have reported on the baseline list and is therefore unobservable. Furthermore, the variance of these quantities ($V[y_{iK}]$) and the covariance between this quantity and the responses to the baseline list ($Cov[y_{i+}^{K-1}, y_{iK}]$) cannot be determined by pre-testing baseline items. However, under certain assumptions we can bound these terms, which will provide a conservative sample size formula.

If we assume that Assumption 1 holds, then y_{iK} is a bernoulli random variable, and due to the properties of covariances, we can write an expression for the necessary sample size based on a desired half-width for a confidence interval (where we assume equal sample sizes within the treatment and baseline groups):

$$n_1^* = (2 \cdot V[y_{i+}^{K-1}] + \sqrt{\pi_K(1-\pi_K)} + 2 \cdot \sqrt{V[y_{i+}^{K-1}]\pi_K(1-\pi_K)} \cdot Corr[y_{i+}^{K-1}, y_{iK}]) \cdot (Z_{\alpha/2}/HW)^2$$

where n_1^* should be rounded up to the nearest integer, and n_2 should be chosen to be at least as large. Therefore, the analyst need only specify the desired half width for the confidence interval (HW), the standard normal quantile corresponding to the desired confidence level ($Z_{\alpha/2}$), the standard deviation of answers to the baseline list (which can be estimated by pre-testing), the anticipated proportion who would honestly say yes to the sensitive item ($\pi_K = .5$ maximizes the sample size), and the anticipated correlation between the sensitive item and the answers to the baseline list ($Corr[y_{i+}^{K-1}, y_{iK}] = 1$ maximizes the sample size).

Appendix B: Choosing the Number of Treatment Units in the List Experiment

With the baseline list chosen so as to minimize the variance of y_{i+}^{K-1}, we can optimally allocate portions of the sample to the baseline and treatment groups. Intuitively, if the variance of the sum of baseline items can be designed to be close to zero, then we can afford to allocate almost all of the sample size to the treatment group because we would only need one baseline observation to establish y_{i+}^{K-1} for all i. It is more reasonable to assume that the baseline list is chosen so as to keep $V[y_{i+}^{K-1}]$ small (and $V[y_{i+}^{K}] - V[y_{i+}^{K-1}] > 0$), and then the design problem simplifies to choosing n_2 (the number of respondents who receive the sensitive item) so as to minimize the variance. We can simplify by treating n_2 as continuous, and since (2) describes a hyperbolic shape, the global minimizing value can be found by setting the first derivative equal to zero and solving the resulting quadratic equation for n_2. I denote this n_2^{opt}, and the solution depends on n, $V[y_{i+}^{K}]$, and $V[y_{i+}^{K}] - V[y_{i+}^{K-1}]$. Under these conditions the real number solution to the quadratic equation that minimizes the variance

of the estimator is the following:

$$n_2^{opt} = \frac{nV[y_{i+}^K]}{V[y_{i+}^K] - V[y_{i+}^{K-1}]} - \frac{\sqrt{(2nV[y_{i+}^K])^2 - 4(V[y_{i+}^K] - V[y_{i+}^{K-1}])V[y_{i+}^K]n^2}}{2(V[y_{i+}^K] - V[y_{i+}^{K-1}])} \tag{4}$$

In practice, n_2^{opt} would have to be rounded to the nearest whole number, and we would need to specify plausible values for $V[y_{i+}^K]$ and $V[y_{i+}^K] - V[y_{i+}^{K-1}]$. However, $V[y_{i+}^{K-1}]$ will be small if the list is designed properly, and if questions from previous surveys are used in the creation of the baseline list, then we can specify plausible values for this number. Furthermore, if $V[y_{i+}^{K-1}]$ is small then $Cov[y_{i+}^{K-1}, y_{iK}]$ will tend to be small, and $V[y_{i+}^K]$ will be dominated by $V[y_{iK}] = \pi_K(1-\pi_K)$. Unfortunately, due to the shape of the objective function, the benefits from unequal sample sizes appear to be minimal for reasonable values of $V[y_{i+}^{K-1}]$.

Appendix C: Design for the Double List Experiment

In the double list experiment (Droitcour et al., 1991), two baseline lists are employed (an A list and a B list), and the sensitive item is randomly included on one of the two lists for each respondent. If we define the following variables:

$$\tilde{y}_{Aik} = \begin{cases} 1 & \text{if individual } i \text{ would honestly say yes to item } k \text{ on the A list} \\ 0 & \text{if individual } i \text{ would honestly say no to item } k \text{ on the A list} \end{cases}$$

$$\tilde{y}_{Bik} = \begin{cases} 1 & \text{if individual } i \text{ would honestly say yes to item } k \text{ on the B list} \\ 0 & \text{if individual } i \text{ would honestly say no to item } k \text{ on the B list} \end{cases}$$

where for simplicity we use the same list size K for both the A and the B lists, and where the sensitive item is the same on both lists, so that $y_{AiK} = y_{BiK}$ for all i. Sums can be defined over these variables:

$$\begin{aligned} \tilde{y}_{Ai+}^K &= \sum_{k=1}^{K} \tilde{y}_{Aik} \qquad & \tilde{y}_{Ai+}^{K-1} &= \sum_{k=1}^{K-1} \tilde{y}_{Aik} \\ \tilde{y}_{Bi+}^K &= \sum_{k=1}^{K} \tilde{y}_{Bik} \qquad & \tilde{y}_{Bi+}^{K-1} &= \sum_{k=1}^{K-1} \tilde{y}_{Bik}. \end{aligned}$$

In contrast to these unobservable opinions, we we observe y_{Ai+}^{K-1} and y_{Bi+}^{K} for the individuals $i = 1, \ldots, n_1$ in the A list baseline group (B list treatment group), and y_{Ai+}^{K} and $y_{Bi+}^{K-1}$for the individuals $i = n_1 + 1, \ldots, n$ in the A list treatment group (B list baseline group).

In the remaining we adapt Assumption 1 for the double list experiment.

Assumption 2 (Honest Responses) $y_{Ai+}^{K-1} = \tilde{y}_{Ai+}^{K-1}$ *and* $y_{Bi+}^{K} = \tilde{y}_{Bi+}^{K}$ *for* $i = 1, ..., n_1$, *and* $y_{Bi+}^{K-1} = \tilde{y}_{Bi+}^{K-1}$ *and* $y_{Ai+}^{K} = \tilde{y}_{Ai+}^{K}$*for* $i = n_1 + 1, \ldots, n$.

Under Assumption 2, we have two unbiased estimators for the parameter of interest,

$$\hat{\pi}_K^A = \frac{1}{n_2} \sum_{i=n_1+1}^{n} y_{Ai+}^{K} - \frac{1}{n_1} \sum_{i=1}^{n_1} y_{Ai+}^{K-1} \tag{5}$$

$$\hat{\pi}_K^B = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{Bi+}^{K} - \frac{1}{n_2} \sum_{i=1+n_1}^{n_2} y_{Bi+}^{K-1}, \tag{6}$$

and these can be averaged to produce an unbiased double list estimator. Notice that the sensitive item must be the same on each list (i.e. $\tilde{y}_{AiK} = \tilde{y}_{BiK}$).

If we assume equal sample sizes ($n_1 = n_2 = n/2$), equal weights to the average of the estimators in 5 and 6, and we define $y_{AiK} \equiv y_{Ai+}^{K} - y_{Ai+}^{K-1}$ and $y_{BiK} \equiv y_{Bi+}^{K} - y_{Bi+}^{K-1}$ to be the implied unobservable differences between the lists, then the double list estimator can be shown to have the following variance:

$$
\begin{aligned}
V[\hat{\pi}_K^A/2 + \hat{\pi}_K^B/2] &= 1/4 \cdot V\left[\frac{1}{n/2}\left(\sum_{i=n/2+1}^{n} y_{Ai+}^K - \sum_{i=1}^{n/2} y_{Ai+}^{K-1}\right) + \frac{1}{n/2}\left(\sum_{i=1}^{n/2} y_{Bi+}^K - \sum_{i=1+n/2}^{n} y_{Bi+}^{K-1}\right)\right] \\
&= 1/4 \cdot V\left[\frac{1}{n/2}\sum_{i=n/2+1}^{n}\left\{y_{Ai+}^K - y_{Bi+}^{K-1}\right\} + \frac{1}{n/2}\sum_{i=1}^{n/2}\left\{y_{Bi+}^K - y_{Ai+}^{K-1}\right\}\right] \\
&= 1/4 \cdot \frac{1}{n/2} V\left[y_{Ai+}^K - y_{Bi+}^{K-1}\right] + 1/4 \cdot \frac{1}{n/2} V\left[y_{Bi+}^K - y_{Ai+}^{K-1}\right] \\
&= \frac{1}{2n}\left\{V[y_{Ai+}^K] + V[y_{Bi+}^{K-1}] - 2 \cdot Cov[y_{Ai+}^K, y_{Bi+}^{K-1}]\right\} \\
&+ \frac{1}{2n}\left\{V[y_{Bi+}^K] + V[y_{Ai+}^{K-1}] - 2 \cdot Cov[y_{Bi+}^K, y_{Ai+}^{K-1}]\right\} \qquad (7) \\
&= \frac{1}{2n}\left\{\left(V[y_{Ai+}^{K-1}] + V[y_{AiK}] + 2 \cdot Cov[y_{Ai+}^{K-1}, y_{AiK}]\right) + V[y_{Bi+}^{K-1}]\right. \\
&- \left. 2 \cdot \left(Cov[y_{Ai+}^{K-1}, y_{Bi+}^{K-1}] + Cov[y_{AiK}, y_{Bi+}^{K-1}]\right)\right\} \\
&+ \frac{1}{2n}\left\{\left(V[y_{Bi+}^{K-1}] + V[y_{BiK}] + 2 \cdot Cov[y_{Bi+}^{K-1}, y_{BiK}]\right) + V[y_{Ai+}^{K-1}]\right. \\
&- \left. 2 \cdot \left(Cov[y_{Bi+}^{K-1}, y_{Ai+}^{K-1}] + Cov[y_{BiK}, y_{Ai+}^{K-1}]\right)\right\} \\
&= \frac{1}{n}\left\{V[y_{Ai+}^{K-1}] + V[y_{Bi+}^{K-1}] + 1/2\left(V[y_{AiK}] + V[y_{BiK}]\right) - 2Cov[y_{Ai+}^{K-1}, y_{Bi+}^{K-1}]\right\} \qquad (8) \\
&+ \frac{1}{n}\left(Cov[y_{AiK}, y_{Ai+}^{K-1}] - Cov[y_{BiK}, y_{Ai+}^{K-1}]\right) \qquad (9) \\
&+ \frac{1}{n}\left(Cov[y_{BiK}, y_{Bi+}^{K-1}] - Cov[y_{AiK}, y_{Bi+}^{K-1}]\right) \qquad (10)
\end{aligned}
$$

where (7) is the variance formula to use in the calculation of standard errors once the double list experiment has been conducted.

Due to the fact that the sensitive items are the same on both lists, $\tilde{y}_{AiK} = \tilde{y}_{BiK}$, when Assumption 2 holds or when we make the weaker assumption that respondents would react the same way to the inclusion of the sensitive item of both lists (i.e. $y_{AiK} = y_{BiK}$), then the (9) and (10) terms are zero, and the variance only includes the terms in (8).

Concentrating on the (8) term, negative correlation design *within* the lists A and B can reduce the terms $V[y_{Ai+}^{K-1}]$ and $V[y_{Bi+}^{K-1}]$, but we can further reduce the variance in (8) by choosing baseline lists A and B so that $Cov[y_{Ai+}^{K-1}, y_{Bi+}^{K-1}]$ is large. Using this formula, and using Assumption 2, we can pre-test the baseline lists to determine possible variances of the double list estimator. The only element of this expression that is unknown is the variance on

the sensitive item ($V[y_{AiK}]$), and plausible values can be plugged in for this quantity. Using these facts, and the properties of covariances, we can write an expression for the necessary sample size based on a desired half-width for a confidence interval (where we assume equal sample sizes within the treatment and baseline groups):

$$n^* = \left\{V[y_{Ai+}^{K-1}] + V[y_{Bi+}^{K-1}] + \pi_K(1-\pi_K) - 2Cov[y_{Ai+}^{K-1}, y_{Bi+}^{K-1}]\right\} \cdot (Z_{\alpha/2}/HW)^2$$

where n^* should be rounded up to the nearest integer. Therefore, if pre-testing has established the aforementioned variances and covariances, the analyst need only specify the desired half width for the confidence interval (HW), the standard normal quantile corresponding to the desired confidence level ($Z_{\alpha/2}$), and the anticipated proportion that would honestly say yes to the sensitive item ($\pi_K = .5$ maximizes the sample size).

Appendix D: The Estimator of Joint Proportions

We define the parameters γ_k^b to be the population proportion that would report k items on the baseline list, and we define γ_k^t to be the population proportion that would report k items on the treatment list for $k = 0, ..., K$, where $\gamma_K^b = 0$ by definition. We denote the observed proportions from the baseline and treatment groups as

$$\hat{\gamma}_k^b = \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{\{y_{i+}^{K-1}=k\}} \text{ for all } k = 0, ..., K-1$$
$$\hat{\gamma}_k^t = \frac{1}{n_2}\sum_{i=n_1+1}^{n_2} 1_{\{y_{i+}^{K}=k\}} \text{ for all } k = 0, ..., K$$

where $1_{\{.\}}$ is an indicator function.

If we further define $\pi_{K|k}$ to be population proportion that, conditional on saying yes to exactly k items on the baseline list, would honestly say yes to the sensitive item, then the the $\boldsymbol{\gamma}^t$ parameters can be written in terms of the $\boldsymbol{\gamma}^b$ and $\boldsymbol{\pi}_{K|k}$ parameters in the following manner (for k=0,...,K):

$$\gamma_k^t = \pi_{K|k-1}\gamma_{k-1}^b + (1-\pi_{K|k})\gamma_k^b \tag{11}$$

where $\pi_{K|-1}$ and $\pi_{K|K}$ are defined to be zero, and $\pi_{K|j}$ is defined to be zero whenever $\gamma_j^b = 0$, so that the joint proportion $\pi_{K|K-1}\gamma_{K-1}^b$ can be written as γ_K^t.

The $\boldsymbol{\gamma}^b$ can also be written (somewhat redundantly) in terms of the $\boldsymbol{\pi}_{K|k}$ parameters (for k=0,...,K-1)

$$\gamma_k^b = \pi_{K|k}\gamma_k^b + (1 - \pi_{K|k})\gamma_k^b \tag{12}$$

Given this notation and the definition of $\pi_{K|K} = \gamma_K^b = 0$, we can now demonstrate that the joint proportions $\pi_{K|k-1}{\cdot}\gamma_{k-1}^b$ can be written in terms of the $\boldsymbol{\gamma}$ parameters (for $k = 1, ..., K-2$)

$$\begin{aligned}
\sum_{j=k}^{K} \gamma_k^t &= \sum_{j=k}^{K} \pi_{K|j-1}\gamma_{j-1}^b + \sum_{j=k}^{K}(1 - \pi_{K|j})\gamma_j^b \\
&= \sum_{j=k-1}^{K-1} \pi_{K|j}\gamma_j^b + \sum_{j=k}^{K-1}(1 - \pi_{K|j})\gamma_j^b \\
-\sum_{j=k}^{K-1} \gamma_k^b &= \sum_{j=k}^{K-1} \pi_{K|j}\gamma_j^b + \sum_{j=k}^{K-1}(1 - \pi_{K|j})\gamma_j^b \\
\hline
&= \pi_{K|k-1}\gamma_{k-1}^b
\end{aligned}$$

Therefore the joint proportions can be estimated with the observed proportions $\hat{\gamma}_k^b$ and $\hat{\gamma}_k^t$ (for k=0,...,K).

Appendix E: Conditional Analysis

Using the result from the previous section, it is straightforward to show that the conditional proportions $\pi_{K|k-1}$ can be estimated using the list experiment data.

$$\begin{aligned}
\pi_{K|k-1}\gamma_{k-1}^b &= \sum_{j=k}^{K} \gamma_k^t - \sum_{j=k}^{K} \gamma_k^b \\
\pi_{K|k-1} &= \frac{\sum_{j=k}^{K} \gamma_k^t - \sum_{j=k}^{K} \gamma_k^b}{\gamma_{k-1}^b} \\
\widehat{\pi}_{K|k-1} &= \frac{\sum_{j=k}^{K} \widehat{\gamma}_k^t - \sum_{j=k}^{K} \widehat{\gamma}_k^b}{\widehat{\gamma}_{k-1}^b}
\end{aligned}$$

Furthermore, notice that the joint population proportion $\pi_{K|k-1}\gamma^b_{k-1}$ can alternatively be interpreted as the proportion that would honestly say yes to the sensitive item, and would say yes to exactly k items on the treatment list, therefore, $\frac{\pi_{K|k-1}\gamma^b_{k-1}}{\gamma^t_k}$ can be interpreted as the population proportion that, conditional on saying yes to exactly k items on the treatment list, would honestly say yes to the sensitive item. This conditional proportion can be estimated in the following manner:

$$\widehat{\pi_{K|k-1}\gamma^b_{k-1}}/\gamma^t_k = \frac{\sum_{j=k}^{K}\widehat{\gamma}^t_k - \sum_{j=k}^{K}\widehat{\gamma}^b_k}{\widehat{\gamma}^t_k}$$

In addition to these conditional proportions, the algorithm for a simple piecewise regression estimator is presented below.

Algorithm for the Piecewise Regression Estimator (with $K-1$ baseline items on):

1. Let $k = 1$
2. Create indicator variables for individuals reporting at least k items on the treatment list.
3. Regress the treatment list indicator variables on the conditioning variables (perhaps with logistic regression).
4. Create indicator variables for individuals reporting at least k items on the baseline list.
5. Regress the baseline list indicator variables on the conditioning variables (perhaps with logistic regression).
6. Form the curve that is the difference between the two regression curves.
7. Repeat for $k = 2, ..., K$
8. Add the K difference curves calculated in Step 6.
9. Truncate at zero or one as necessary.

Appendix F: The Piecewise Estimator and the MLE

We can write the difference-in-means list experiment estimator as the following:

$$\hat{\pi}_K = \sum_{j=1}^{K}\left[\sum_{k=j}^{K}\hat{\gamma}^t_k - \sum_{k=j}^{K-1}\hat{\gamma}^b_k\right] \tag{13}$$

where the second term in the brackets is defined to be zero when $j > K-1$. The piecewise estimator constrains estimates of $\pi_{K|k}\gamma_k^b$ and $\pi_{K|k}$ to be in the zero-one range ($\widehat{\gamma}_k^b$ will be within the admissible range automatically),

$$\hat{\pi}_K^{piece} = \sum_{j=1}^{K}\left[\min\{\hat{\gamma}_{k-1}^b, \max\{0, \sum_{k=j}^{K}\hat{\gamma}_k^t - \sum_{k=j}^{K-1}\hat{\gamma}_k^b\}\}\right] \tag{14}$$

In contrast, the maximum likelihood estimator calculates the values of the parameters that maximize the following log-likelihood function.

$$\begin{aligned} l = & \sum_{i=1}^{n_1}\left\{\sum_{k=0}^{K-1} 1_{y_{i+}^{K-1}=k}\log\gamma_k^b\right\} \\ & + \sum_{i=n_1+1}^{n}\Big\{1_{y_{i+}^K=0}\log\gamma_0^b(1-\pi_{K|0}) \\ & + \sum_{k=1}^{K-1}[1_{y_{i+}^K=k}\log(\gamma_{k-1}^b\pi_{K|k-1} + \gamma_k^b(1-\pi_{K|k}))] \\ & + 1_{y_{i+}^K=K}\log\gamma_{K-1}\pi_{K|K-1}\Big\} \end{aligned}$$

Notice that the maximum likelihood estimator imposes the same constraints on the $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ parameters as the piecewise estimator. However, unlike the piecewise estimator, the MLE simultaneously estimates all the parameters. In particular, the piecewise estimator only uses the baseline data to estimate the $\boldsymbol{\gamma}$ parameters, while the MLE uses both the baseline and treatment data to estimate the $\boldsymbol{\gamma}$ parameters.

Appendix G: Simple Test to Diagnose Underreporting

If we assume that baseline group respondents honestly report their answers to the baseline list, and we assume that treatment group respondents may underreport their answers to the treatment list, but will not overreport their answers, then we can test for underreporting in the treatment group. The formal statement of the assumption follows:

Assumption 3 (Baseline Honest Responses, Treatment Never Overreport) $y_{i+}^{K-1} = \tilde{y}_{i+}^{K-1}$ *for* $i = 1, ..., n_1$, *and* $y_{i+}^K \leq \tilde{y}_{i+}^K$ *for* $i = n_1+1, \ldots, n$.

In this situation, a simple test for underreporting on the treatment list would involve testing whether the joint proportion $\pi_{K|K-2}\gamma^b_{K-2}$ is significantly different from zero in the negative direction. This parameter can be written solely in terms of the $\boldsymbol{\gamma}$ parameters in the following manner:

$$\begin{aligned}(\gamma^t_K + \gamma^t_{K-1}) - \gamma^b_{K-1} &= \pi_{K|K-1}\gamma^b_{K-1} + (1-\pi_{K|K-1})\gamma^b_{K-1} + \pi_{K|K-2}\gamma^b_{K-2} \\ &\quad - \pi_{K|K-1}\gamma^b_{K-1} + (1-\pi_{K|K-1})\gamma^b_{K-1} \\ &= \pi_{K|K-2}\gamma^b_{K-2}\end{aligned}$$

If the test statistic $(\widehat{\gamma}^t_K + \widehat{\gamma}^t_{K-1}) - \widehat{\gamma}^b_{K-1}$ can be shown to be significantly less than zero, then when Assumption 3 holds, this can be taken as evidence of underreporting. Because this test statistic is a difference between two independent proportions, the test can be conducted in the standard manner.

References

Berinsky, A.J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43(4):1209–1230.

Berinsky, A.J., and A.N. Glynn. 2010. "Design for the List Experiment." *Working Paper* .

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

Coutts, Elisabeth, and Ben Jann. 2009. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." *Unpublished Manuscript* .

Dalton, D.R., J.C. Wimbush, and C.M. Daily. 1994. "Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior." *Personnel Psychology* 47(4):817–828.

Droitcour, J., R.A. Caspar, M.L. Hubbard, T.L. Parsley, W. Visscher, and T.M. Ezzati. 1991. "The item count technique as a method of indirect questioning: A review of its development and a case study application." *Measurement Error in Surveys. New York: Wiley* .

Glynn, A.N., D.J. Greiner, and K.M. Quinn. 2010. "Social Desirability and the Self Reported Vote for Obama." *Working Paper* .

Gonzalez-Octanos, E., C. Kiewiet de Jonge, C. Melendez, J. Osorio, and D.W. Nickerson. 2010. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *Working Paper* .

Hubbard, M.L., R.A. Casper, J.T. Lessler, et al. 1989. "Respondent reactions to item count lists and randomized response." *Proceedings of the Survey Research Section of the American Statistical Association* pp. 544–548.

Imai, K. 2010. "Statistical Analysis of the Item Count Technique." *Working Paper* .

Kane, J.G., S.C. Craig, and K.D. Wald. 2004. "Religion and Presidential Politics in Florida: A List Experiment*." *Social Science Quarterly* 85(2):281–293.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(01):49–69.

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

Kuklinski, J.H., P.M. Sniderman, K. Knight, T. Piazza, P.E. Tetlock, G.R. Lawrence, and B. Mellers. 1997b. "Racial Prejudice and Attitudes Toward Affirmative Action." *American Journal of Political Science* 41(2):402–419.

Miller, J. 1984. "A New Survey Technique for Studying Deviant Behavior." *Ph.D. Dissertation, Sociology Department, The George Washington University* .

Raghavarao, D., and WT Federer. 1979. "Block total response as an alternative to the randomized response method in surveys." *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1):40–45.

Silver, B.D., B.A. Anderson, and P.R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80(2):613–624.

Sniderman, P.M., and E.G. Carmines. 1997. *Reaching beyond race*. Harvard Univ Pr.

Sniderman, P.M., P.E. Tetlock, and T. Piazza. 1992. "Codebook for the 1991 National Race and Politics Survey."

Streb, M.J., B. Burrell, B. Frederick, and M.A. Genovese. 2008. "Social Desirability Effects and Support for a Female American President." *Public Opinion Quarterly* 72(1):76.

Tsuchiya, T. 2005. "Domain estimators for the item count technique." *Survey Methodology* 31:41–51.

Tsuchiya, T., Y. Hirai, and S. Ono. 2007. "A Study of the Properties of the Item Count Technique." *Public Opinion Quarterly* 71(2):253.

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