Public Opinion Quarterly, Vol. 77, Special Issue, 2013, pp. 159–172

WHAT CAN WE LEARN WITH STATISTICAL TRUTH SERUM?
DESIGN AND ANALYSIS OF THE LIST EXPERIMENT

ADAM N. GLYNN*

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 reduce dishonest or evasive responses. However, standard practice with the list experiment requires a large sample size, utilizes only a difference-in-means estimator, and does not provide a measure of the sensitive item for each respondent. This paper addresses all of these issues. First, the paper presents design principles for the standard list experiment (and the double list experiment) for the reduction of bias and variance as well as providing sample-size formulas for the planning of studies. Second, this paper proves that a respondent-level probabilistic measure for the sensitive item can be derived. This provides a basis for diagnostics, improved estimation, and regression analysis. The techniques in this paper are illustrated with a list experiment from the 2008–2009 American National Election Studies (ANES) Panel Study and an adaptation of this experiment.

Introduction

The sensitivity of some topics presents measurement problems for many areas of social science research. Sex, drugs, crime, religion, race, and politics are all inherently sensitive subjects. Furthermore, the inaccuracy of measurement due to the use of sensitive survey questions is not trivial. In a canonical example,

ADAM GLYNN is an associate professor of government at Harvard University, Cambridge, MA, USA. An earlier version of this article was presented at the 2010 Midwest Political Science Association conference and at the 2010 Summer Conference for the Society of Political Methodology. The author would like to thank the editors and three anonymous reviewers as well as Adam Berinsky, Matt Blackwell, Justin Grimmer, Chase Harrison, Sunshine Hillygus, Kosuke Imai, Gary King, Richard Nielsen, Dave Peterson, Kevin Quinn, and Dustin Tingley for their helpful comments and suggestions. The usual caveat applies. *Address correspondence to Adam Glynn, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA; e-mail: aglynn@fas.harvard.edu.

doi:10.1093/poq/nfs070

studies have shown that routinely a quarter or more of respondents who report voting actually did not vote (e.g., see Silver, Anderson, and Abramson 1986). The effect of sensitive questions on our inferences may become more problematic once we consider the inclusion of "don't know" answers or the effects of item nonresponse (Berinsky 1999).

One solution to this problem has been the use of aggregation techniques, where respondents are 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, Wimbush, and Daily 1994), or the list experiment (Sniderman and Carmines 1997; Kuklinski, Cobb, and Gilens 1997 Kuklinski et al. 1997).

Recently, the list experiment has gained in popularity.[1] To some extent, this is due to the relative ease of administration and the apparent success of the technique in some applications. Dalton, Wimbush, and Daily (1994), LaBrie and Earleywine (2000), and Tsuchiya, Hirai, and Ono (2007) all found that the list experiment provided increased estimates of socially undesirable behaviors in comparison to direct questioning and provided statistically similar estimates for nonsensitive items. However, in a survey of studies, Tourangeau and Yan (2007) find the efficacy of the technique to be variable.

Although the list experiment provides an alternative to direct questioning, at least two major difficulties with the list experiment have limited its efficacy and use. First, the list experiment usually requires a large sample size in order to achieve reasonable levels of precision (Tsuchiya 2005; Tsuchiya, Hirai, and Ono 2007; Corstange 2009). Second, standard analysis of the list experiment does not provide a measure of the sensitive item for each respondent. Such a measure would sometimes allow the detection of dishonest reporting and would also allow the use of the list experiment in a multivariate regression framework.[2]

This paper makes four contributions in solving these problems. First, it presents design principles for the reduction of bias and variance in the single list

1. For examples, see Ahart and Sackett (2004); Biemer and Brown (2005); Blair and Imai (2012); Brueckner, Morning, and Nelson (2005); Comsa and Postelnicu (2012); Coutts and Jann (2011); Diaz-Cayeros et al. (2011); Flavin and Keane (2009); Gilens, Sniderman, and Kuklinski (1998); Heerwig and McCabe (2009); Janus (2010); Martinez and Craig (2010); Rayburn, Earleywine, and Davison (2003); Redlawsk, Tolbert, and Franko (2010); Walsh and Braithwaite (2008); Wimbush and Dalton (1997); Hubbard et al. (1989); Holbrook and Krosnick (2010a, 2010b); and Gonzalez-Octanos et al. (2012).

2. In groundbreaking 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.

experiment. Second, in the context of the double list experiment (Droitcour et al. 1991), it presents additional design principles for the reduction of bias. Third, sample-size calculations are provided that allow the analyst to plan future studies on the basis of pretested baseline lists. Finally, it is demonstrated that respondent-level probabilistic measures of the sensitive item can be identified from the list experiment. These quantities have recently been used to provide alternative estimators, multivariate analysis, and advanced diagnostics for the list experiment (Imai 2011; Blair and Imai 2012).

The Standard List Experiment

The list experiment works by aggregating the sensitive item with a list of nonsensitive items. For example, on the 2008–2009 American National Election Studies (ANES) Panel Study, one set of respondents was randomized to a baseline (or control) group and was asked the following question:

> Below are four things. Please tell us how many of them you would dislike. We do not need to know which ones you would dislike, just how many.
>
> 1. *Listening to music*
> 2. *Making it legal for two men to marry*
> 3. *Getting a phone call from a telemarketer*
> 4. *Being a garbage collector*

Another set of respondents was randomized to a *treatment* group to receive the same list with the following sensitive item appended (although the list order was randomized):

> 5. *A black person becoming president*

There are three things to note about this design.[3] First, all respondents are provided with some level of privacy protection in terms of the sensitive item. Baseline group respondents do not receive the sensitive item within the list experiment, and treatment group respondents who answer with anything less than "five" could deny disliking a black president.

Second, if respondents in both groups answer the question honestly, then the randomization into baseline and treatment groups allows the analyst to estimate the proportion for the sensitive item 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,

3. On the ANES, there was also a third group that received the item "Barack Obama becoming president" instead of the item "A black person becoming president." This group is not relevant for the discussion of this paper.

Table 1. Estimated Mean Level of Dislike for a Black President of the United States among White Respondents on the 2008–2009 ANES Panel Study List Experiment (standard errors in parentheses)

	Treatment list (four items and black president item)	Baseline list (four items)	Treatment-minus-baseline difference
Estimates	2.234	2.230	0.005
	(0.044)	(0.038)	(0.058)
n	431	405	

NOTE.—Numbers were rounded after calculations were made.
SOURCE.—2008–2009 American National Election Studies Panel Study

and all of the other survey error components are negligible, then this estimator will be unbiased for the proportion in the population that would dislike a black person becoming president (Droitcour et al. 1991). 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. The difference-in-means data from the ANES for white respondents is presented in table 1. The estimated proportion of white respondents who dislike a black president is statistically indistinguishable from zero.

Third, this design leads to estimates with high variance when compared to a direct question because the sensitive item is aggregated with a number of nonsensitive items. To some extent, this additional variance is the cost of the perceived privacy protection; however, the remainder of this paper discusses design innovations to alleviate this difficulty.

Design and Analysis for the Single List Experiment

PREVIOUS DESIGN ADVICE

The use of the list experiment relies on the assumption that respondents answer the list questions honestly. Therefore, most list experiment designers have focused on standards that give respondents the privacy necessary to allow for honest responses. Most importantly, "ceiling effects" (Kuklinski, Cobb, and Gilens 1997) can occur when a respondent would honestly respond "yes" to all nonsensitive items. When this happens, a treatment group respondent no longer has the protection to honestly report their response to the sensitive item. Furthermore, near-ceiling effects may also be possible. For example, respondents who would report three nonsensitive items (with a four-item baseline list) may underreport the number of items on the treatment list because they do not want to show the possibility 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, the use of many high-prevalence nonsensitive items, which would increase the occurrence of ceiling effects, should be avoided (Droitcour et al. 1991). Second, the use of many low-prevalence nonsensitive items should be avoided. If respondents are aware that all the nonsensitive items have low prevalence, they may become concerned about the level of privacy protection and underreport their answers (Tsuchiya, Hirai, and Ono 2007). In other words, list designers are told to avoid too many low-variance items (high prevalence or low prevalence).[4] Third, lists should not be too short because short lists will also tend to increase the likelihood of ceiling effects (Kuklinski, Cobb, and Gilens 1997).

Unfortunately, these three pieces of design advice tend to lead to increased variance (Tsuchiya, Hirai, and Ono 2007; Corstange 2009).[5] This state of affairs presents the list designer with an unfortunate trade-off between ceiling effects and variable results (an apparent bias-variance trade-off). However, it may be possible to reduce both bias and variance by constructing items that are negatively correlated.

REDUCING BIAS AND VARIANCE WITH NEGATIVE CORRELATION

As an example of list experiment design, consider the ANES baseline list presented above. The results from this list experiment are presented in table 1. This list successfully minimizes the variance of responses (standard deviation of $0.76 = 0.038 \bullet \sqrt{405}$);[6] however, it uses many low-variance items to accomplish this goal: one low prevalence (listening to music) and two high

4. 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, Cobb, and Gilens (1997) warn against contrast effects (where the resonance of the sensitive item overwhelms the baseline items), and Droitcour et al. (1991) advise that the nonsensitive 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 nonsensitive items, and to choose baseline items that do not seem out of place with the sensitive item. Sometimes, an attempt at coherence between the sensitive and baseline items may lead to concerns that the baseline items are themselves sensitive. For example, the "gay marriage" item in the ANES list might be seen as sensitive by some respondents.

5. In addition to increased variance, Tsuchiya, Hirai, and Ono (2007) present evidence that longer lists may produce measurement error due to the inability of respondents to remember their responses to all items on the list.

6. The standard deviation was 0.78 among all ANES baseline respondents. This appears to be a reduction in standard deviation when compared to the standard four-item list adapted by Kane, Craig, and Wald (2004) from the canonical three-item list developed for the 1991 National Race and Politics Survey (Sniderman, Tetlock, and Piazza 1992) and used in Kuklinski, Cobb, and Gilens (1997) and Kuklinski et al. (1997). Although the comparison is not exact because the results come from different surveys and the question wording was not the same ("dislike" versus "angry"), the results from Streb et al. (2008), table 1 (which uses the Kane, Craig, and Wald [2004] baseline list), imply a standard deviation of 0.96 (although this could be as high as 1.09 or as low as 0.80 due to rounding).

prevalence (getting a phone call from a telemarketer, being a garbage collector). We might prefer to maintain (or reduce) the variance of responses while reducing the number of low-variance items.

The following baseline list (A) accomplishes both of these goals by inducing negative correlation within the list:

Baseline List A

1. *Listening to music*
2. *Making it legal for two men to marry*
3. *Teaching intelligent design along with evolution in public schools*
4. *Getting a phone call from a telemarketer*

Notice that the only change from the ANES list is that the "Being a garbage collector" item has been replaced with "Teaching intelligent design along with evolution in public schools." The "intelligent design" item is designed to have greater variance than the "garbage" item. However, it is also designed to have a negative correlation with the "gay marriage" item (respondents who dislike gay marriage are less likely to dislike the teaching of intelligent design, so few respondents will dislike both items).

In order to demonstrate the potential of the technique, a list experiment using baseline list A was fielded on a Mechanical Turk Internet survey (Berinsky, Lenz, and Huber 2012) conducted December 1–8, 2010, and reported in table 2 (screenshots are available in the supplementary data's online appendix A).[7] This survey used a convenience sample, so the results should not be generalized to any population; however, note that despite removing one of the low-variance items from the ANES list, the induced negative correlation has led to a reduction in the variance of the baseline responses. The standard error and the sample size for the baseline group in this survey imply a standard deviation of baseline responses of $0.65 = 0.39 \bullet \sqrt{280}$, which is smaller than the standard deviation of 0.76 from the ANES list.

While the baseline list can be pretested without fielding the sensitive item, the variance of the estimator is also a function of the variance of the sensitive item and the covariation between the sensitive item and the baseline list.[8]

7. These data were collected in conjunction with a December 2010 Political Experiments Research Lab (PERL) Study (Adam Berinsky, PI). This was a convenience sample; however, we note that respondents from Mechanical Turk samples tend to be more similar to the U.S. population than in-person convenience samples (Berinsky, Lenz, and Huber 2012). There were two respondents in each treatment group who answered the List A question but did not answer the List B question discussed later in the context of the double list experiment. These have been removed from the analysis in order to simplify the discussion. Their inclusion does not affect the results appreciably.

8. This covariation is likely to be minimized by minimizing the variance of the baseline list. If there are two baseline lists with the same variance, then the designer might try to anticipate and choose the list that minimizes the covariation. However, if the designer has two good baseline lists, then the double list experiment is an option (discussed in the next section).

Table 2. Estimated Mean Level of Dislike for a Black President of the United States among White Respondents from the Baseline List A List Experiment (standard errors in parentheses)

	Treatment list (four items and black president item)	Baseline list (four items)	Treatment-minus-baseline difference
Estimates	1.761	1.732	0.029
	(0.046)	(0.039)	(0.060)
n	272	280	

NOTE.—Numbers were rounded after calculations were made.
SOURCE.—December 2010 Mechanical Turk Internet Survey

However, even without pretesting on these components, we can perform conservative sample-size calculations. If we allocate half of the sample to each treatment group, and we assume maximum variance for the sensitive item and the correlation between the baseline list and the sensitive item, and if we write the variance of the baseline list as V, then the sample size needed to guarantee a $(1 - \alpha)\%$ confidence interval with a half-width no greater than HW is

$$n = (z_{a/2} / HW)^2 (4 \bullet V + 1/2 + 2 \bullet \sqrt{V}), \quad (1)$$

where $z_{a/2}$ is the critical value from the normal distribution (see the supplementary data's online appendix B for details).

ESTIMATING JOINT AND CONDITIONAL PROPORTIONS FOR DIAGNOSTICS

Although 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. As demonstrated below, the piecewise table of responses to the treatment and baseline lists can be used to estimate joint and conditional proportions. These estimated proportions can be used for diagnostic tests of the assumption of honest responses when estimates are significantly less than zero or greater than one.

Table 3 reports the data from the baseline list A experiment. The first and third rows report the response proportions from the list experiment using baseline list A. The first row shows the proportions for the "treatment" respondents who received the sensitive item appended to baseline list A, and the third row shows the proportions for the "baseline/control" respondents. For the treatment and control lists and for each possible number of indicated items, we estimate the proportion reporting at least that number (rows 2 and 4). The fifth row of table 3 presents the difference between rows 2 and 4 and can be interpreted as estimates of joint proportions that

Table 3. Estimated Piecewise Proportions on the Treatment and Baseline/Control Lists among White Respondents from the Baseline List A List Experiment

		Number disliked						
		0	1	2	3	4	5	Sum
1	Treatment	0.022	0.346	0.504	0.107	0.022	0.000	1.000
2	Treatment "at least"	1.000	0.978	0.632	0.129	0.022	0.000	
3	Control	0.014	0.343	0.539	0.104	0.000	0.000	1.000
4	Control "at least"	1.000	0.986	0.643	0.104	0.000	0.000	
2 – 4	Joint	0.000	–0.008	–0.011	0.025	0.022	0.000	0.029
$\frac{2-4}{1}$	Conditional	0.000	–0.022	–0.021	0.235	1.000	n/a	

NOTE.—Rows 1 and 3 present the estimated proportions disliking each particular number of items on the treatment and control lists. Rows 2 and 4 present the estimated proportions disliking at least a particular number of items on the treatment and control lists. Row 5 presents the estimated difference between rows 2 and 4, which is an estimate of the proportion that dislike a black president and the total number of treatment list items indicated by the column. Row 6 presents the ratio of rows 5 and 1, which is an estimate of the proportion that dislike a president among those that report disliking the total number of treatment list items indicated by the column. Numbers were rounded after calculations were made, and negative estimates of proportions have been retained for illustrative purposes.

SOURCE.—December 2010 Mechanical Turk Internet Survey (272 respondents in the treatment group and 280 respondents in the control group)

dislike the number of treatment list items indicated by the column label and also dislike the sensitive item (online appendix C provides a proof; see the supplementary data online). Note that the sum of these entries reconstructs the difference-in-means estimate from table 2.

The sixth row of table 3 divides the fifth row by row 1 and can be interpreted as the conditional probability of disliking the sensitive item conditional on disliking the number of treatment list items indicated by the column labels. The row 6 estimates can also be interpreted as respondent-level probabilistic measures of the sensitive item, and along with row 5, these form the basis of a test of behavioral assumptions (because proportions must be between zero and one).

For the joint proportions (row 5), it is possible for estimates to be less than zero. In table 3, this occurred twice (–0.008 and –0.011). These negative estimates could be due to violations of the behavioral assumptions or could be due to chance, but because the joint proportion estimates are the difference between proportions from independent samples, we can use a standard test of whether the row 2 proportions are significantly smaller than the row 4 proportions. For example, note that the estimated proportion disliking at least two items on the treatment list is 0.632 (based on 272 observations) and this is less than the estimated proportion disliking at least two items on the control list, 0.643 (based on 280 control list observations).

The one-sided test produces a p-value of 0.43, so we cannot rule out the possibility that the negative estimate of –0.011 is due to chance. We also cannot rule out the possibility that the –0.008 estimate is due to chance.

For the conditional proportions (row 6), it is possible to get estimates greater than 1. This did not occur in table 3; however, in the case that it did, we would want to test whether these proportions were due to chance. This can be accomplished by testing whether the row 2 minus row 1 proportion (i.e., the proportion in the treatment group that disliked "more than" the column label) is greater than the row 4 proportion. Again, this is a standard test based on the difference in proportions.

In table 3, we did not find a row 5 estimate that was significantly less than zero or a row 6 estimate that was significantly greater than one. However, if one finds an estimate significantly less than zero or greater than one, we would need to utilize the more advanced diagnostic techniques presented in Blair and Imai (2012) in order to alleviate the multiple testing implicit in this procedure. Additionally, the joint proportions can be used for alternative population-level estimators and regression estimators (see the supplementary data's online appendix D; see also Imai [2011] and Blair and Imai [2012]).

Design and Analysis for the Double List Experiment

Although the negative correlation between pairs of baseline items can help minimize the bias and variance, 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, the sample-size calculation in equation (1) 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. The Double List Experiment (DLE), introduced in Droitcour et al. (1991), alleviates these deficiencies.

The DLE uses two baseline lists. For example, we combine baseline list A with the following list, denoted as baseline list B:

Baseline List B:

1. *Watching movies*
2. *Making it legal for two men to form a civil union*
3. *Teaching creationism along with evolution in public schools*
4. *Being a garbage collector*

In the DLE, the respondents are again separated randomly into two groups, but both groups function simultaneously as baseline and treatment groups. For example, in the aforementioned Internet survey, a group of 280 white respondents first received baseline list A and then received baseline list B with

Table 4. Estimated Mean Level of Dislike for a Black President of the United States among White Respondents from the Double List Experiment (standard errors in parentheses)

	Treatment list (four items and black president item)	Baseline list (four items)	Treatment-minus-baseline difference
Estimates	1.761	1.732	0.029
	(0.046)	(0.039)	(0.060)
Estimates	1.400	1.301	0.099
	(0.050)	(0.048)	(0.070)
Mean estimate			0.064
			(0.033)
n	272	280	

NOTE.—Numbers were rounded after calculations were made.
SOURCE.—December 2010 Mechanical Turk Internet Survey

the sensitive item appended. The 272 white respondents in the other group received baseline list A with the sensitive item appended and then received baseline list B. In this way, the first group of respondents functions as the baseline group with respect to baseline list A and the treatment group with respect to baseline list B, while the second group functions as the baseline group with respect to baseline list B and the treatment group with respect to baseline list A. Therefore, we now have two list experiments, and both provide difference-in-means estimators that can be averaged (see table 4). Droitcour et al. (1991) provide a variance formula for the average of these estimators, demonstrating a reduction in variance when compared to the single list experiment. However, there are additional design opportunities available for the DLE.

The negative correlation design from the previous section can be applied to both baseline lists in the DLE, but we can also reduce variance by using two baseline lists that have a positive correlation on the responses to the "how many" question (see the supplementary data's online appendix E 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 the sensitive question. Positive correlation between the lists is an attempt to approach this level of precision.

As an example, consider that the second item from baseline list A, "Making it legal for two men to marry," is similar to the first item on baseline list B, "Making it legal for two men to form a civil union," 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 a positive correlation will be induced between the responses to the two baseline lists. In fact, due to positive correlation between the lists, the average of the two difference-in-means from

this DLE produced a standard error of 3.3 percent with the 552 respondents (see table 4). This represents almost a 50-percent reduction in the standard error, although the improvement might be greater with a different baseline list B. Baseline list B presented here underperformed compared to baseline list A, producing a single list standard error of 7 percent.

Perhaps more importantly, the DLE allows for more precise planning if the baseline items for both lists are pretested. 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. The variance of the estimator will be minimized when 1) negative correlation within each baseline list minimizes the variance of each baseline list; and 2) positive correlation between each list maximizes the covariance between the lists (see the supplementary data's online appendix E for a proof). If we write the variance of baseline list A as V^A, the variance of baseline list B as V^B, and the covariance between the baseline lists as Cov^{AB}, then with the aforementioned assumptions, the sample size needed to guarantee a $(1-\alpha)\%$ confidence interval with a half-width no greater than HW is

$$n = (z_{a/2} / HW)^2 (V^A + V^B + 1/4 - 2 \bullet Cov^{AB}), \tag{2}$$

where again $z_{a/2}$ is the critical value from the normal distribution (see the supplementary data's online appendix E for the derivation).

Although the DLE allows a reduction in variance when compared to the single list experiment, the use of the technique raises some additional concerns. First, all of the potential design and implementation problems for the list experiment are worsened. The possibility of poor question wording and the length of the interview will increase, as well as the cognitive load on the respondent. Second, the positive correlation between the lists could increase the cognitive load of the task. Therefore, pretesting the baseline lists will be critical. Third, if the sensitive item is included on a list for all respondents, this may preclude directly asking the sensitive item. This can be avoided by leaving the sensitive item off of both lists for a randomly selected subset of the respondents. Finally, the use of two lists raises the concern of list-order effects and the possibility that the first list will affect the second list.

The diagnostics of the previous section can be used separately on the A and B list experiments, but there are additional diagnostic opportunities within the DLE. A partial check can be conducted by testing whether the results from the A and B list experiments are significantly different. This test can be conducted by bootstrapping respondents. Additionally, the sample can be split in half so that the first half of the sample receives the DLE with the A list first

and the second half of the sample receives the DLE with the B list first. This provides the opportunity to assess list-order effects for the DLE. Finally, if effects appear to be different across the different list experiments, the second list experiment can be discarded without invalidating the first experiment.

Conclusion

The list experiment represents an increasingly popular tool for indirectly asking sensitive questions on surveys. However, list experiments tend to require large sample sizes, and standard difference-in-means estimators do not provide respondent-level measures, multivariate analysis, or diagnostics.

This paper makes four contributions in solving these problems. First, the design principle of negative within-list correlation reduces variance and bias due to ceiling effects, without using high- or low-prevalence items. Second, the design principle of positive between-list correlation reduces variance beyond the standard reduction associated with the double list experiment of Droitcour et al. (1991). Third, the sample-size calculations allow the analyst to plan future studies on the basis of pretested baseline lists. Finally, the joint and conditional proportions in the piecewise table provide the basis for diagnostics, alternative estimators, multivariate regression analysis, and respondent-level measurements. Imai (2011) has recently shown that these quantities can be used for multivariate regression analysis with a maximum likelihood estimator that reduces root mean square error. Blair and Imai (2012) have recently shown that these quantities can be used to provide extensive diagnostics for the list experiment.

However, whereas this paper has demonstrated the potential benefits from taking design and analysis seriously within the single and double list experiments, there are undoubtedly other related aggregated response designs that might further reduce bias and variance. One that has already received some attention is the LISTIT procedure introduced in Corstange (2009) and reanalyzed in Blair and Imai (2012). Another is the inclusion of an additional low-prevalence item on the baseline list in order to equalize the lengths of the baseline and treatment lists (Cobb 2001). Future research should focus on developing designs and procedures for further improvements. Blair and Imai (2012) represent an important first step in this respect.

Supplementary Data

Supplementary data are freely available online at http://poq.oxfordjournals.org.

References

Ahart, Allison M., and Paul R. Sackett. 2004. "A New Method of Examining Relationships between Individual Difference Measures and Sensitive Behavior Criteria: Evaluating the Unmatched-Count Technique." *Organizational Research Methods* 7:101–14.

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43:1209–30.

Berinsky, Adam J., Gabriel S. Lenz, and Gregory A. Huber. 2012. "Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research." *Political Analysis* 20:351–68.

Biemer, Paul, and Gordon Brown. 2005. "Model-Based Estimation of Drug-Use Prevalence Using Item-Count Data." *Journal of Official Statistics* 21:287–308.

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

Brueckner, Hannah, Ann Morning, and Alondra Nelson. 2005. "The Expression of Biological Concepts of Race." Paper presented at the Annual Meeting of the American Sociological Association, Philadelphia, PA, USA.

Cobb, Michael. 2001. "Unobtrusively Measuring Racial Attitudes: The Consequences of Social Disability Effects." Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Comsa, Mircea, and Camil Postelnicu. 2012. "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item-Count Technique." *International Journal of Public Opinion Research*.

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

Coutts, Elisabeth, and Ben Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched-Count Technique (UCT)." *Sociological Methods & Research* 40:169–93.

Dalton, Dan R., James C. Wimbush, and Catherine M. Daily. 1994. "Using the Unmatched-Count Technique (UCT) to Estimate Base Rates for Sensitive Behavior." *Personnel Psychology* 47:817–28.

Diaz-Cayeros, Alberto, Beatriz Magaloni, Aila Matanock, and Vidal Romero. 2011. "Living in Fear: Social Penetration of Criminal Organizations in Mexico." Center for U.S.-Mexican Studies, University of California, San Diego.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 1991. "The Item-Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application." In *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A.Mathiowetz, and S. Sudman, 185–210. New York: John Wiley & Sons.

Flavin, Patrick, and Michael Keane. 2009. "How Angry Am I? Let Me Count the Ways: Question Format Bias in List Experiments." Department of Political Science, University of Notre Dame.

Gilens, Martin, Paul M. Sniderman, and James H. Kuklinski. 1998. "Affirmative Action and the Politics of Realignment." *British Journal of Political Science* 28:159–83.

Gonzalez-Octanos, Ezequiel, Chad Kiewiet de Jonge, Carlos Melendez, Javier Osorio, and David W. Nickerson. 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56:20–217.

Heerwig, Jennifer A., and Brian J. McCabe. 2009. "Education and Social Desirability Bias: The Case of a Black Presidential Candidate." *Social Science Quarterly* 90:674–86.

Holbrook, Allison L., and Jon A. Krosnick. 2010a. "Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling into Question the Method's Validity." *Public Opinion Quarterly* 74:328–43.

———. 2010b. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item-Count Technique." *Public Opinion Quarterly* 74:37–67.

Hubbard, Michael L., Rachel A. Casper, and Judith T. Lessler. 1989. "Respondent Reactions to Item-Count Lists and Randomized Response." Proceedings of the Survey Research Section of the American Statistical Association, 544–48.

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

Janus, Alexander L. 2010. "The Influence of Social Desirability Pressures on Expressed Immigration Attitudes." *Social Science Quarterly* 91:928–46.

Kane, James G., Stephen C. Craig, and Kenneth D. Wald. 2004. "Religion and Presidential Politics in Florida: A List Experiment." *Social Science Quarterly* 85:281–93.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the 'New South.'" *Journal of Politics* 59:323–49.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997. "Racial Prejudice and Attitudes toward Affirmative Action." *American Journal of Political Science* 41:402–19.

LaBrie, Joseph W., and Mitchell Earleywine. 2000. "Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique." *Journal of Sex Research* 37:321–26.

Martinez, Michael D., and Stephen C. Craig. 2010. "Race and 2008 Presidential Politics in Florida: A List Experiment." *Forum* 8:Article 4.

Miller, Judith. 1984. "A New Survey Technique for Studying Deviant Behavior." Ph.D. dissertation, George Washington University.

Raghavarao, Damaraju, and Walter T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society Series B (Methodological)* 41:40–45.

Rayburn, Nadine Recker, Mitchell Earleywine, and Gerald C. Davison. 2003. "An Investigation of Base Rates of Anti-Gay Hate Crimes Using the Unmatched-Count Technique." *Journal of Aggression, Maltreatment & Trauma* 6:137–52.

Redlawsk, David P., Caroline J. Tolbert, and William Franko. 2010. "Voters, Emotions, and Race in 2008: Obama as the First Black President." *Political Research Quarterly* 63:875–89.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80:613–24.

Sniderman, Paul M., and Edward G. Carmines. 1997. *Reaching beyond Race*. Cambridge, MA: Harvard University Press.

Sniderman, Paul M., Philip E. Tetlock, and Thomas Piazza. 1992. "Codebook for the 1991 National Race and Politics Survey." Berkeley, CA: Survey Research Center. Available at http://sda.berkeley.edu/D3/Natlrace/Doc/nrac.htm.

Streb, Matthew J., Barbara Burrell, Brian Frederick, and Michael A. Genovese. 2008. "Social Desirability Effects and Support for a Female American President." *Public Opinion Quarterly* 72:76–89.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–83.

Tsuchiya, Takahiro. 2005. "Domain Estimators for the Item-Count Technique." *Survey Methodology* 31:41–51.

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

Walsh, Jeffrey A., and Jeremy Braithwaite. 2008. "Self-Reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-Count Technique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students." *Journal of Alcohol and Drug Education* 52:49–74.

Wimbush, James C., and Dan R. Dalton. 1997. "Base Rate for Employee Theft: Convergence of Multiple Methods." *Journal of Applied Psychology* 82:756–63.