Public Opinion Quarterly

(MIS)MEASURING SENSITIVE ATTITUDES WITH THE LIST EXPERIMENT
SOLUTIONS TO LIST EXPERIMENT BREAKDOWN IN KENYA

ERIC KRAMON
KEITH WEGHORST*

Abstract List experiments (LEs) are an increasingly popular survey research tool for measuring sensitive attitudes and behaviors. However, there is evidence that list experiments sometimes produce unreasonable estimates. Why do list experiments "fail," and how can the performance of the list experiment be improved? Using evidence from Kenya, we hypothesize that the length and complexity of the LE format make them costlier for respondents to complete and thus prone to comprehension and reporting errors. First, we show that list experiments encounter difficulties with simple, nonsensitive lists about food consumption and daily activities: over 40 percent of respondents provide inconsistent responses between list experiment and direct question formats. These errors are concentrated among less numerate and less educated respondents, offering evidence that the errors are driven by the complexity and difficulty of list experiments. Second, we examine list experiments measuring attitudes about political violence. The standard list experiment reveals lower rates of support for political violence compared to simply asking directly about this sensitive attitude, which we interpret as list experiment breakdown. We evaluate two modifications to the list experiment designed to reduce its complexity: private tabulation and cartoon visual aids. Both modifications greatly enhance list experiment performance, especially among respondent subgroups where the standard procedure is most problematic. The paper makes two key contributions: (1) showing that techniques such as the list experiment, which have promise for reducing response bias, can introduce different forms of error associated with question complexity and difficulty; and (2) demonstrating the effectiveness of easy-to-implement solutions to the problem.

Eric Kramon is an assistant professor of political science and international affairs at George Washington University, Washington, DC, USA. KEITH WEGHORST is an assistant professor of political science at Vanderbilt University, Nashville, TN, USA. *Address correspondence to Keith Weghorst, Department of Political Science, 230 Appleton Place, PMB 0505, Nashville, TN 37212, USA; email: keith.weghorst@vanderbilt.edu.

doi:10.1093/poq/nfz009

Survey researchers are often concerned with measuring sensitive attitudes and behaviors, including support for political violence, experience with corruption, and racial attitudes. A major challenge for studying such topics with surveys is social desirability bias: many individuals do not want to reveal socially unacceptable or potentially illegal attitudes and behaviors. Scholars have developed a number of strategies for reducing sensitivity-driven measurement error. The list experiment—or "item count technique"—is one approach that is increasingly popular in political science and related disciplines.

In this paper, we evaluate two modifications to standard list experiment procedures. The first allows respondents to privately tabulate the number of items in the list that apply, thereby aiding accurate response while creating additional assurance of privacy. The second modification adds visual aids, which is intended to reduce respondent error—particularly among respondents who find the instructions and demands of a list experiment challenging.

List experiments (LEs) reduce survey error by asking respondents about sensitive issues indirectly: sensitive items are embedded in a list with several nonsensitive items, and participants are asked how many items they agree with or apply to them, but not which ones (see examples found in tables 3 and 4 later in this paper). This approach reduces the perceived costs/risks of answering honestly. However, enthusiasm surrounding the list experiment has drawn attention away from its potential limitations. The length and complexity of the question format make them prone to comprehension and reporting errors. Importantly, such errors may be concentrated among certain population subgroups—those without experience answering complex survey questions or those who most prevalently hold the sensitive attitude of interest. Unfortunately, identifying the extent to which these issues bias list experimental data is challenging because survey respondents' "true" answers to sensitive questions are usually unknown (Simpser 2017). Nonetheless, LEs often break down in obvious ways: producing estimates that are lower than the direct question, or even nonsensical ones, such as negative estimates (Holbrook and Krosnick 2010). In that light, we are motivated by two questions: Why do list experiments sometimes "fail" or break down? How can the performance of the list experiment be improved?

In this paper, we examine the LE and its ability to reduce survey error in Kenya, where we sought to measure public support for political violence. First, we investigate the performance of the LE using lists of simple, nonsensitive items about food consumption and daily activities. We show that the LE encounters difficulties with these simple and nonsensitive lists: over 40 percent of respondents provide inconsistent responses in LE versus direct question formats. These "failures" are concentrated among less numerate and less educated respondents, evidence that errors are driven by LE question complexity and difficulty.

Second, we turn to list experiments designed to measure attitudes about political violence. We find that the standard LE estimates *lower* rates of

support for political violence than those obtained by asking directly. These underestimates are most pronounced among less educated participants and those who provided inconsistent responses in the nonsensitive LEs described above, evidence that technique difficulty is driving list experiment breakdown.

Finally, we evaluate two low-cost, context-appropriate modifications to the list experiment designed to reduce the complexity of the technique. The first allows for private tabulation, and the second combines private tabulation with cartoon visual aids. We find that both modifications improve list experiment performance, including among the subgroups that had difficulty with the nonsensitive LE.

This paper contributes to the literature on survey response bias in two ways. First, we show that indirect techniques such as the list experiment, which have promise for reducing response bias, can introduce different forms of error that are associated with question complexity and difficulty. This is important because the survey literature is populated with list experiments that perform well; we highlight limitations that might not be obvious from reading this published literature because of publication bias and the "file drawer problem." Our aim is not to suggest that all LEs are problematic, but rather to draw attention to these limitations.

Our second contribution is demonstrating that relatively easy-to-implement and low-cost modifications can greatly enhance the performance of the technique, especially among populations where the standard procedure is most problematic. Modifications designed to reduce item complexity and difficulty can be adapted by applied survey researchers working in a range of contexts.

Measuring Sensitive Attitudes with the List Experiment

Attitudes toward violence are emblematic of the challenges of studying sensitive topics. Support for political violence is subject to under-reporting biases because such violence is illegal and its approval is generally socially undesirable. Past research on violence has addressed sensitivity-driven measurement error by alleviating perceived costs/risks of answering truthfully. Strategies include asking about violent behavior indirectly (Humphreys and Weinstein 2006), administering sensitive survey modules separately from a larger survey (Scacco 2016), anticipating or controlling for enumerator ethnicity effects (Kasara 2013; Carlson 2014; Adida et al. 2016), or one of several experimental approaches: endorsement experiments (Blair et al. 2013; Lyall, Blair, and Imai 2013), randomized response technique (Blair, Imai, and Zhou 2015), or the list experiment.

The list experiment is a promising alternative to direct questions, offering respondents greater secrecy for sensitive responses (e.g., Kuklinski, Cobb, and Gilens 1997; Corstange 2018; Gonzalez-Ocantos et al. 2011; Blair and Imai 2012; Glynn 2013). The LE presents a sensitive statement as one of many items of a list and asks respondents to identify how many total list items apply to them. Participants are randomly assigned to either a treatment list including the

sensitive item or a control list that does not. Because the lists are otherwise identical, and assignment is randomized, the difference in means between treatment and control lists can be attributed to the sensitive item. If successfully implemented, the technique yields an estimate of the prevalence of the sensitive attitude.

Two assumptions must be satisfied for LE estimates to be valid: "no-liars" and "no design effects" (Blair and Imai 2012). The first states that respondents "do not lie about the sensitive item" (Rosenfeld, Imai, and Shapiro 2016, 795). The second requires that adding the sensitive item to a list does not change the way respondents engage with control items. Lists are generally designed to avoid "floor" and "ceiling" effects, which undermine how the sensitive attitude is rendered undetectable (Glynn 2013).

For single LEs, the estimated prevalence of the sensitive item is the difference-in-means between treatment and control groups (e.g., Blair and Imai 2012; Streb et al. 2008). For example, if the control group mean is 2 and the treatment group mean is 2.2, the estimate in the sample would be 20 percent. In the double list experiment design (DLE), which uses two sets of lists such that all respondents receive one control list and one treatment list, the estimate is a weighted average of the two single list experiment estimates (Glynn 2013). We discuss the DLE further in Section "The standard list experiment".

DO LIST EXPERIMENTS WORK?

Demonstrating that a list experiment "works" is difficult because the true prevalence of a given sensitive item is unknown (Simpser 2017). To our knowledge, there exists only one study of LE performance with a sensitive item where objective validation is possible; Rosenfeld, Imai, and Shapiro (2016) find that the LE reduces response bias, but less efficiently than other indirect alternatives. Efficiency concerns are echoed in other meta-analyses (Blair, Coppock, and Moor 2018).

One overview of 48 studies with list experiments finds that only 63 percent of them "worked." LEs in the remaining studies were unsuccessful—either revealing *lower prevalence* of the sensitive attitude compared to direct questions or statistically indistinguishable estimates (Holbrook and Krosnick 2010). Such issues span sensitive topics, including drug and alcohol use (Droitcour et al. 1991; LaBrie and Earleywine 2000; Biemer and Brown 2005), shoplifting (Tsuchiya, Hirai, and Ono 2007), citizenship activities (Prior 2009), intergroup prejudice (Kane, Craig, and Wald 2004), same-sex marriage (Lax, Phillips, and Stollwerk 2016), and theft (Ahart and Scakett 2004).

List experiment breakdowns may predominate in certain population subgroups (Zigerelli 2011), which can be especially damaging for inferences. These findings raise important concerns about the LE and echo worry over a "file-drawer" problem whereby most LE failures are never reported (Gelman 2014).

WHY DO LIST EXPERIMENTS BREAK DOWN?

Most list experiment research has focused on how the technique reduces perceived risks of divulging a socially undesirable attitude. We focus on a different and fundamental issue: LEs are more complex and difficult than direct questions and require more effort to answer honestly. This can introduce unanticipated measurement error that, unlike downward sensitivity biases, operates less predictably and may inflate or deflate estimates (de Jonge and Nickerson 2014).

Answering survey questions is demanding. To complete a question "optimally," respondents must interpret question meaning, comb their memories for relevant information, aggregate information into general views, and report these views with precision (Tourangeau 1984; Krosnick 1991). List experiments are complex and demand more effort to optimally answer than conventional questions for several reasons. First, LE instructions are more extensive and less familiar than most question prompts. Thus, even understanding LE procedures requires more effort than conventional questions. Second, list experiments often use control items that probe about experiences or attitudes requiring longer-term recall. Third, respondents must perform the higher-order task of individually considering a set of qualitative statements and the arithmetic task of adding up items on a list. As a result, LEs introduce an "honesty-effort compromise," whereby they reduce perceived costs of being honest about the sensitive item but require additional effort to actually answer truthfully.

We argue that these features can lead to list experiment breakdown, particularly in low-development settings where the technique is increasingly implemented. In such environments, complex survey design may be unfamiliar, skills the LE requires may not be used on a day-to-day basis, and willingness to expend the additional effort required to complete them may vary. We anticipate that these obstacles will be more common among less educated and less numerate respondents, for whom the LE will require the most exertion. Thus, LE failures may be concentrated among these groups.

Research Design

We conducted this research in Kenya in June 2012, one year before the 2013 general elections. We fielded the survey in four areas within the capital of Nairobi—Githurai, Karangware, Kibera, and Mathare. To varying degrees, each site experienced election-related violence and unrest following Kenya's disputed elections in 2007. We designed the survey with two goals: (1) to identify list experiment breakdowns and the subgroups where they are most prevalent; and (2) to test modifications to improve list experiment performance through LEs about attitudes toward political violence.

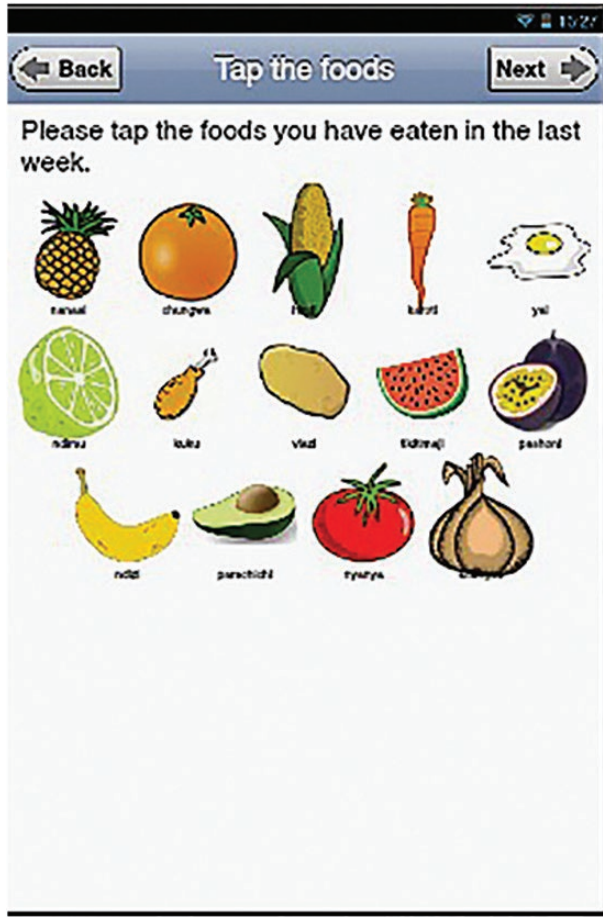

Figure 1. Nonsensitive foods direct questions.

In what follows, we describe our strategies for achieving both objectives. We then detail survey design (summarized in figure 4), highlighting how it allows us to make inferences about the performance of our design modifications. We conclude this section by discussing sampling and sample characteristics. Further survey procedure details are found in Online Appendix B and Online Appendix C.

DETECTING FAILURES USING NONSENSITIVE LIST EXPERIMENTS

We first assess whether list experiments fail due to the additional effort they require to report honestly. We do so by comparing responses to direct questions with responses to the same items in LE format. Breakdowns are observable when responses to direct questions do not match the numerical response

provided in an LE. For example, if a respondent replies "3" to the LE question, they should indicate that three of the corresponding statements apply when asked directly. In standard list experiments, sensitivity biases make it difficult to interpret mismatches between LE data and direct responses. In addition, when LE items are complex or require substantial thought, responses to list and direct items may differ as an artifact of the question format (e.g., Flavin and Keane 2009; de Jonge and Nickerson 2014). Detecting list experiment "failures" is therefore challenging with existing LE data.

We designed two list experiments with the explicit goal of assessing LE failures directly attributable to technique difficulty: a food consumption LE and a daily activities LE. At the beginning of the survey, respondents received general instructions regarding the list experiment. The goal was to increase familiarity with the format from the outset and to prepare respondents for the questions ahead. They were given the following prompt:

> *For most of the rest of the survey, we're going to ask questions in a way you may not be familiar with. We will read you a list of many items and ask you how many of them apply. For this, you will give a number but you will not tell me which items you agree with. So if I ask you a number of activities you do every day and one of the activities is going to the market, if you go to the market—I don't want you to tell me. I want to you count that as one thing you do every day and then add up in your head the total number.*

Table 1 presents the exact question wording and list items. These items are entirely nonsensitive so that direct questions about them should not be subject to response bias. They are also simple in order to make recall easy and minimize response variability driven by question format (Flavin and Keane 2009).

The first nonsensitive list asked participants about foods consumed in the prior week. The particular foods were selected based on existing data on consumption patterns in Kenya (see Online Appendix D). In the second nonsensitive list, we presented respondents with a number of everyday activities. Some activities capture counting and enumerating, like sending currency via

Table 1. Nonsensitive list experiments

Food consumption list	Daily activities list
How many of the following foods have you eaten in the last week?	How many of the following things have you done in the past month?
Banana	Read a newspaper
Chicken	Loaned money to a friend
Avocado	Sent money using Mpesa
Egg	Sold a product for profit
Tomato	Deposited money at a bank

mobile phone ("Mpesa") or depositing money into a bank. The activities are commonplace, but not so universal as to be subject to over-reporting biases. For example, 96 percent of Kenyan households have at least one Mpesa account (Suri and Jack 2016) and two-thirds of Nairobians read a newspaper at least monthly (Afrobarometer Round 5). In addition, 24 percent of Kenyans use bank services and 15 percent informally borrow from shopkeepers, friends, and family (Central Bank of Kenya, Kenya National Bureau of Statistics, and FSD Kenya 2016). Usage rates of formal, mobile financial services are similar across wealth quintiles, education, gender, and age (Central Bank of Kenya, Kenya National Bureau of Statistics, and FSD Kenya 2016).

After the LEs, enumerators distracted respondents with a demographic question. They then administered direct questions about food consumption and everyday activities. To simplify the food questions, respondents viewed pictures of 14 total food items with their Swahili name underneath, including the five from the list experiment. This is shown in figure 1. Interviewees used tablets to tap each food they had eaten in the previous week, and the tablet recorded which items they selected (enumerators read instructions in *Sheng*).

Direct questions on daily activities were posed orally (see table 2). We asked participants whether they had engaged in each of the five activities in the LE and included four additional items requiring numeracy skill and experience. We use these nine activities to construct a numeracy metric.

We measure list experiment failure in two ways. First, we calculate the difference between the list response and the direct item response. We count the number of foods that the participant tapped (that were also included in the food LE) and subtract this number from the numerical response provided in the food LE. Similarly, we total the number of activities a respondent reports directly and subtract that from the numeric response given in the daily activities LE. We interpret deviations from 0 as LE breakdown. Second, we create a dichotomous indicator of *list experiment failure* that takes a value of 1 if the continuous measure does not equal 0 and 0 otherwise.[1]

TESTING LIST EXPERIMENT MODIFICATIONS BY MEASURING ATTITUDES ABOUT POLITICAL VIOLENCE

Our second goal is to test the performance of two modifications to the list experiment. The modifications were designed to reduce complexity, especially the difficulties that low-education, low-numeracy respondents may experience. Our analysis of the modifications serves two purposes. First, if the LE performs better when the modifications are used, this provides evidence that the LE is failing at least in part because of its difficulty. Second,

1. One concern is that people may be more likely to respond "yes" to questions posed directly. If so, the continuous measure would have a skewed distribution indicative of over-reporting the direct item. The distribution is relatively symmetric (figure 5) and respondents both under- and over-report LE responses compared to direct questions.

Table 2. Direct questions for daily activities

I am going to read you a list of activities. Please indicate the activities you have done during the past month.
Read a newspaper
Served a meal for guests
Sent money using Mpesa
Sold a product for profit
Loaned a friend money
Deposited money at a bank
Bargained to buy something for a discount
Repaid a debt
Bought vegetables at the market

the modifications provide easy-to-implement solutions to the challenges we identity above.

We test the performance of the two modifications when implementing a list experiment that measures attitudes about the justifiability of political violence. This is an especially sensitive topic in Kenya given its experience with post-election violence. The direct survey item is the following:

> *Please tell me whether you agree or disagree with the following statement: If another tribe steals an election, it is justifiable to use violence to stop them.*
> [Response options are Agree, Disagree, (Not read aloud: Neutral/Non Response)][2]

We chose this wording deliberately. Kenyan elections regularly feature violence, and support for such violence is stigmatized. At the same time, it is plausible that some respondents would agree with this statement; that is, while we do not expect many Kenyans to support political violence generally, some may believe that it is justified under certain conditions. Indeed, the violence surrounding Kenya's 2007 elections, which directly impacted 25 percent of the population and resulted in over 1,000 deaths (Finkel, Horowitz, and Rojo-Mendoza 2012), broke out amid suspicions that the incumbent had rigged the election, denying victory to the main opposition candidate, Raila Odinga. Our analyses will test list experiment performance using responses to this direct question as a lower-bound estimate of justifiability of political violence, assuming the topic's sensitivity will bias direct responses in only one direction: downward. Our sensitive item is thus appropriate for the research context: support for election violence is a socially undesirable attitude in Kenya but may be viewed by some as justifiable under some conditions.

2. We understand the term "tribe" has negative connotations in many settings. In Kenya, however, people speak of "tribalism" and "tribes," not ethnic groups and ethnic identities.

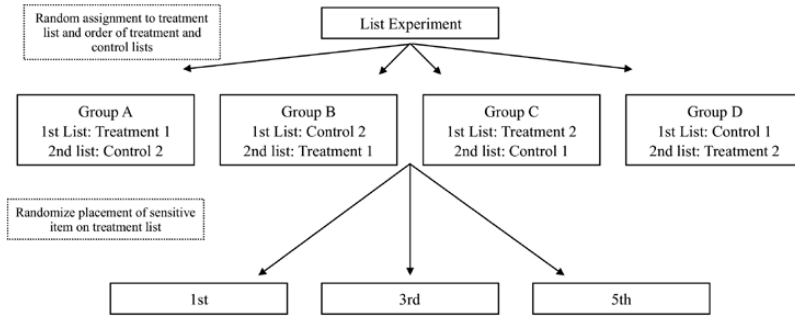

Figure 2. Double list experiment summary.

The Standard List Experiment

All participants completed the standard version of the list experiment. To increase statistical power, we implement double list experiments (DLEs). In the DLE, we use two lists of nonsensitive items (1 and 2). Each participant receives one treatment list *and* one control list: for example, those who receive the treatment item with list 1 (treatment list 1) receive the control list from list 2 (control group 2). To control for potential ordering effects, we randomize the order in which each participant receives the treatment and control lists. In each LE, there are thus four groups: (A) Treatment list 1 + Control list 2; (B) Control list 2 + Treatment list 1; (C) Treatment list 2 + Control list 1; and (D) Control list 1+ Treatment list 2. To control sensitive item location within the treatment lists, we randomize whether the sensitive item is listed in the first, third, or fifth position. Figure 2 visually presents the double list experiment procedure.

Table 3 displays the items for lists 1 and 2. The nonsensitive items were adapted from Afrobarometer questions with known response distributions in order to minimize risk of design effects. Each list included two negatively correlated items in order to reduce potential bias and variance of list experiment estimates (Glynn 2013) and two low-variance items (Tourangeau and Yan 2007).[3] We conducted preliminary field tests of the modifications (described below) in Zanzibar (Tanzania), a setting where elections also feature political violence. That study validated the reliability of Swahili language versions of the nonsensitive list items and corresponding modifications. Finally, the nonsensitive lists were vetted with the field enumeration team prior to implementation in order to adjust for context and adapt formal Swahili to the dialect most widely spoken in Nairobi.

Table 4 presents the instructions for enumerators and the question prompt for the standard procedure. In most LEs administered face-to-face outside the

3. The Afrobarometer questions we draw from present respondents with opposing statements and ask which is more applicable. These resulting list items oppose each other and probe nonsensitive political attitudes. This strategy reduces variance and alleviates potential "no design effects" violations (Glynn 2013).

Table 3. Double list experiment items

List 1
- The government should close news stations that write lies.
- Other Kenyans are not at all trustworthy.
- In line with our customs, we should respect our elders.
- News stations should be free to write whatever they want.
- [*If another tribe tries to steal an election, it is justified to use violence to try to stop them.**]

List 2
- We should NOT have term limits for the president.
- It is better if all children go to school, even if there are not enough books.
- Matatu drivers should do driving exams from time to time.
- It is better if all children in school have books, even if there are not enough books for all children to attend school.
- [*If another tribe tries to steal an election, it is justified to use violence to try to stop them.**]

* Sensitive item included only when respondents received Treatment List 1 or Treatment List 2. The position of the sensitive item is randomized to be in the first, third, or fifth position on the list.

United States, the enumerator either reads the list items out loud, or provides a handout with the list items (Corstange 2009; Gonzalez-Ocantos et al. 2011; Lyall, Blair, and Imai 2013; Oliveros 2016; Frye et al. 2017).[4] Since literacy is a concern in contexts like Kenya, we implemented the standard procedure by having the enumerator read the items out loud, an approach used by others in similar contexts (Corstange 2009).

In the double list design, those who receive treatment list 1 (along with control list 2) serve as the control group for those who receive treatment list 2. Those who receive treatment list 2 (along with control list 1) serve as the control group for those who receive treatment list 1. To produce an aggregated LE estimate, we calculate the weighted average of the two LEs (Glynn 2013). Standard errors are calculated using the variance formula for the DLE in Droitcour et al. (1991).

The Modifications

List handout and private tabulation. The first modification provided participants a textual aid and allowed them to privately tabulate their answer. Respondents were handed a laminated sheet of paper including the lists to which they were assigned and a dry-erase marker. The enumerator instructed interviewees to use the marker on the laminated sheet to help them count up the items. Respondents were shown how to remove dry erase from the laminated sheet, so it was clear their notes would be undetectable. After a respondent

4. Notably, de Cao and Lutz (2018) cleverly ask respondents to hide their hands and transfer stones they are provided from one hand to another for each applicable item, easing aggregation.

Table 4. Instructions and prompts for the sensitive list experiments

Standard procedure:
READ THE FOLLOWING. DO NOT SHOW THE LIST TO THE RESPONDENT: *Now I am going to read you a list of statements. Please tell me how many of these statements are true for you. Please do not tell me which ones, just how many. If you need me to repeat the list, I will. Please do not tell me which ones.*
Private tabulation:
HAND RESPONDENT SHEET [corresponding to assigned list]. AFTER YOU HAND THE SHEET, READ THE FOLLOWING: *Please read each of the following statements. While you read them I will turn my back. Use the pen to put a check mark next to the statements that you agree with. Then count the check marks. Before I turn back around, erase the check marks and tell me how many of the items you checked. Do not tell me which ones, just how many.*
Cartoon:
HAND RESPONDENT SHEET [corresponding to assigned list]. AFTER YOU HAND THE SHEET, READ THE FOLLOWING: *Each picture represents a statement. Please tell me how many of the statements you agree with. Do not tell me how many, just which ones. I will read each statement while you look at the pictures.*

NOTE.—Table presents the instructions for enumerators and question prompts for each of the modified list experiment procedures.

understood this, the enumerator turned 90 degrees away from the respondent, read LE instructions, and implemented the list experiment. Upon completing the question, the enumerator returned 90 degrees to face the respondent and collected the materials.[5]

We designed the private tabulation procedure with the core goal of reducing effort required for respondents to report truthful list experiment responses. By allowing respondents to privately tick along each applicable item, the procedure more closely resembled the effort required to answer several yes/no questions and eased the burden of aggregation. By having enumerators physically turning

5. We note that enumerators turned away from respondents while they completed the modified list experiment, providing additional privacy. Our modifications also are thus a bundled treatment that simultaneously reduces question format complexity and enhances privacy. If our modifications enhance LE performance among certain respondent subgroups, we are confident these successes can be attributed to reduced complexity, rather than added privacy, which should impact all respondents evenly. Further, research has found that additional privacy provisions and reminders might actually induce survey bias (Singer, Hippler, and Schwarz 1992).

from respondents, our modifications also enhanced privacy. Literacy is required, meaning some population subgroups will benefit less from this modification.

Cartoon handout. Our second modification provides respondents with cartoon visual aids. To ensure that they were contextually appropriate, we worked with a local cartoonist to create an illustration corresponding with each list item. Each cartoon was placed on a handout and numbered by its ordering in the LE (e.g., the first list item was labeled "1"). After providing the respondent with the handout, the enumerator read the script for each list. To ensure comprehension, the enumerator also read each item on the list while the participant looked at the visual aids. Enumerators turned 90 degrees from the respondent to maximize the cartoon modification's comparability to the private tabulation procedure.

Like the tabulation procedure, our goal was to reduce the complexity of the question format and the effort required to answer it. This approach also further eased aggregation by allowing respondents to count the pictures corresponding with their attitudes, rather than reviewing written statements. Essentially, this procedure replicates the private tabulation modification but with visual images rather than text. Table 4 includes the instructions.

Figure 3 provides an illustration of the sensitive item. This cartoon was produced by a Kenyan artist, with the text *mwizi* ("thief") and *matokeo ya urais* ("presidential results") added later to enhance message clarity. Importantly, accusations of election fraud are commonly advanced against both incumbents

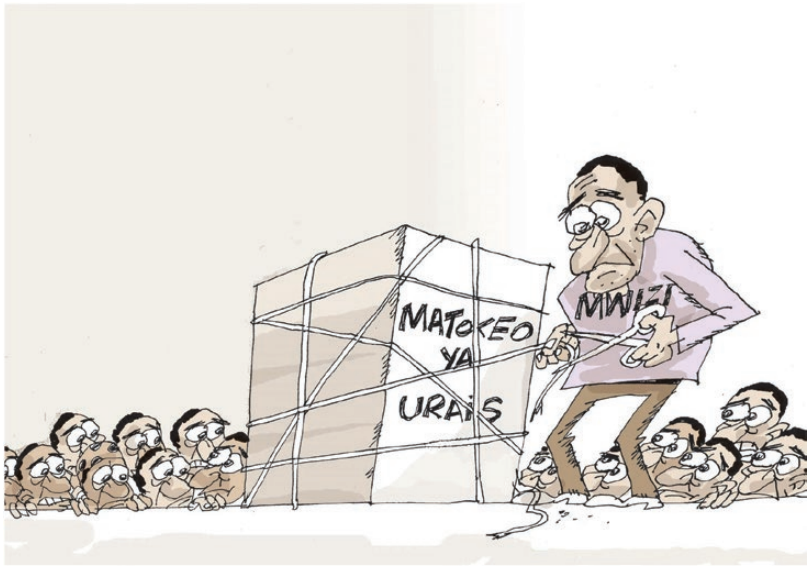

Figure 3. Cartoon corresponding with sensitive item.

and opposition sides and validated by external observers (EU-EOM 2007), suggesting that neither the sensitive item nor the corresponding image are subject to partisan bias.

Online Appendix E shows the cartoons corresponding with nonsensitive items on the lists.

SURVEY DESIGN

We designed the survey to permit within-subjects comparisons of the performance of our two modifications with the LE's standard implementation. Each respondent participated in the same LE at two different points in the survey. All 478 participants completed the standard list experiment, and each participant completed one of the two modifications. We randomly assigned 253 to the cartoon modification and 225 to private tabulation. Randomization was implemented at the individual level using a random-number generator in the survey software. Figure 4 summarizes the sequence of the survey and the research design.

As each respondent participated in list experiments that were identical except in the manner that they were implemented, we distracted participants after the first LE with unrelated questions. This included the food and activities component described earlier (see figure 4). We also asked the direct version of the sensitive question on violence after all of the LEs were completed. We did so to avoid priming respondents before they completed the list experiments.[6]

Importantly, we randomized the order in which participants completed the standard versus the modified list experiments. As illustrated in figure 4, participants were randomly assigned to one of four groups, each with a different sequence. Randomizing order controls for potential ordering or priming effects.[7] It also ensures that the LE format is not confounded with location in the survey. This is important because survey respondents may provide higher-quality responses at different points in a survey due to fatigue.

6. Posing the direct question after list experiments might cue respondents into our study's goal, but this does not have obvious bias implications for results. On the one hand, it might further downward bias direct reports due to additional sensitivity. On the other hand, respondents could inflate direct question reports to mask LE responses, even though it risks divulging the sensitive attitude. List experiment failures are symmetric, suggesting neither is at play. We asked the direct form of the sensitive item among a set of several direct questions, further reducing its conspicuousness.

7. This includes ordering effects due to performing an LE under two different sets of instructions or learning across completing list experiments. Generally, such effects should bias against finding differences between the standard and modified LEs.

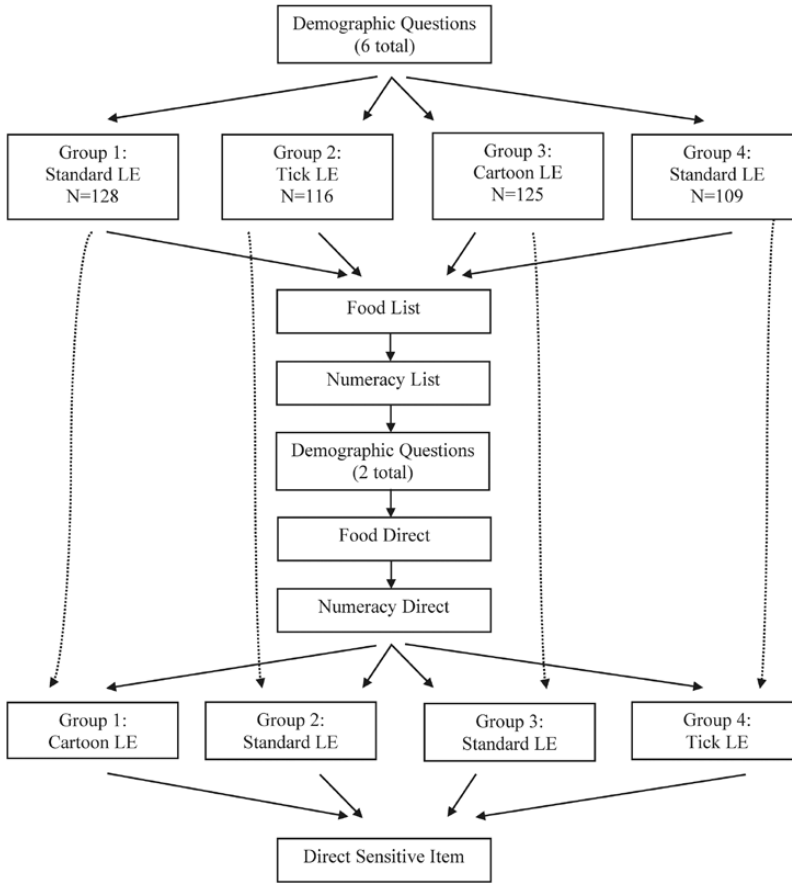

Figure 4. Survey sequence.

SAMPLING AND SAMPLE CHARACTERISTICS

The survey sample includes 478 subjects. Respondents were recruited by survey enumerators through face-to-face contact in four research sites in Nairobi (Kenya's capital): Githurai, Karangware, Kibera, and Mathare. These study sites are dense, mixed residential and commercial areas. The sample is a convenience sample drawn from the population of individuals who live or work within these neighborhoods. The areas feature a combination of permanent structures and informal, temporary residences and businesses, and are thus characteristic of urban areas in Africa and elsewhere.

Respondents were recruited spontaneously by enumerators, who introduced themselves, described the project, and completed the informed consent process. Inclusion in our convenience sample is thus based on willingness or availability of respondents and enumerators did not record refusal rates. While

this approach poses no problems for internal validity—our central goal is testing performance of the design modifications and not to estimate population parameters—we must be cautious about generalizability.

Table A.1 in Online Appendix A provides descriptive statistics and compares our sample to the Afrobarometer Round 5 sample (conducted in 2011). Our sample is more male and slightly younger than the overall population of Nairobi. It has similar proportions of people with no formal education or whose highest level of attainment is primary schooling, a higher percentage who completed secondary school, and a lower percentage who have at least some postsecondary education. Our sample is comparable to the Afrobarometer sample regarding how frequently participants have gone without a cash income in the previous year. Nairobi is an urban area, so our sample is younger, more educated, and less likely to have gone without a cash income than the full population of Kenya.

Our sample thus differs in some ways from the population of Nairobi and Kenya. We are therefore cautious about generalizing beyond our sample. We note, however, that this may present less of a challenge to the generalizability of the subgroup analyses. Although not necessarily representative of Nairobi, we have no reason to believe that the subgroups of less educated or less numerate participants are any different from the larger population of such individuals in Nairobi.

Online Appendix A shows that random assignment distributed respondents evenly into each of the four potential list experiment orderings and each double list experiment suborder. We also controlled for design effects by randomizing the position of the sensitive item as the first, third, or fifth item, and respondents were evenly distributed into these three groups.[8] Respondents in the randomly assigned groups are roughly comparable on a range of observable covariates, including gender, age, and education.

Results

DOES THE NONSENSITIVE TOPIC LIST EXPERIMENT WORK?

We test for list experiment failures by comparing responses given to direct questions about nonsensitive activities to those same items in the LE format. Figure 5 presents the distribution of the continuous measure capturing the difference between the list and direct item responses. The *x*-axis shows the difference between the number of foods/activities reported through the list experiment and the total reported from individually posed direct questions. A value of 0 indicates that responses for direct and list items matched.

8. To permit direct comparison between our modifications and the standard procedure, each individual receives the same treatment placement for both of their LEs. Participants thus participated in LEs that were exactly the same, except for the mode of implementation.

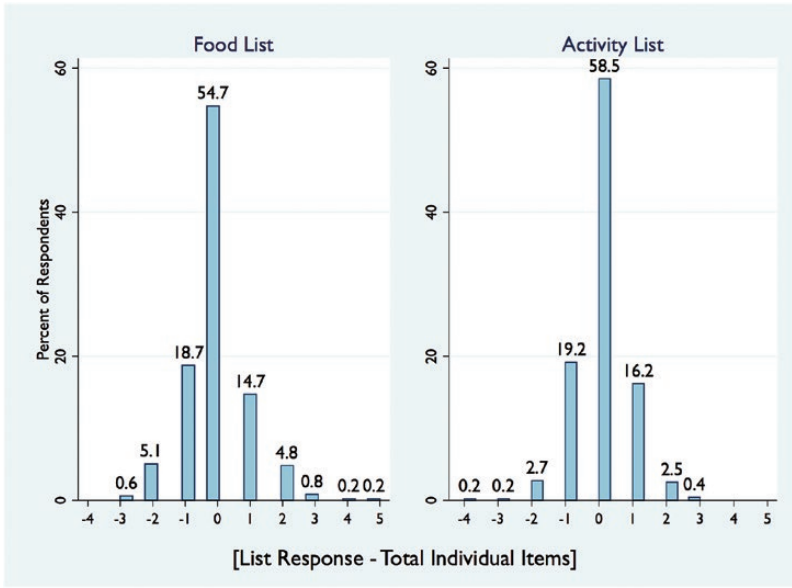

Figure 5. Comparing estimates from the list experiments and direct question. Figure displays the distribution of the continuous measure of list experiment failure. The variable is calculated by adding up the number of positive ("yes") responses to the direct food/activities questions and subtracting that number from the numeric response provided in the list experiment.

Strikingly, less than 60 percent of respondents provide consistent information elicited through these two formats. The mismatch between the two question formats operates in both upward and downward directions, meaning that respondents both over- and under-report behaviors when elicited through the LE design.[9] The symmetrical pattern of failures follows theoretical expectations regarding "non-strategic" respondent error for list experiments (Ahlquist 2018; Blair, Choy, and Imai, forthcoming).

Over 40 percent of our sample provides different responses to LE versus direct questions about innocuous aspects of their day-to-day lives. The number of inconsistent responses may even be higher, as a proportion of those who matched across the two question formats may have offered an "accurate" answer due to chance. This suggests that a large proportion of respondents may not be optimally answering list experiment questions: considering each item and accurately aggregating them into a numeric response.

9. Roughly 10 percent of respondents reported all foods/activities. While analogous to a "ceiling effect," no "design effects" violations occur because we do not use these lists to study a sensitive attitude.

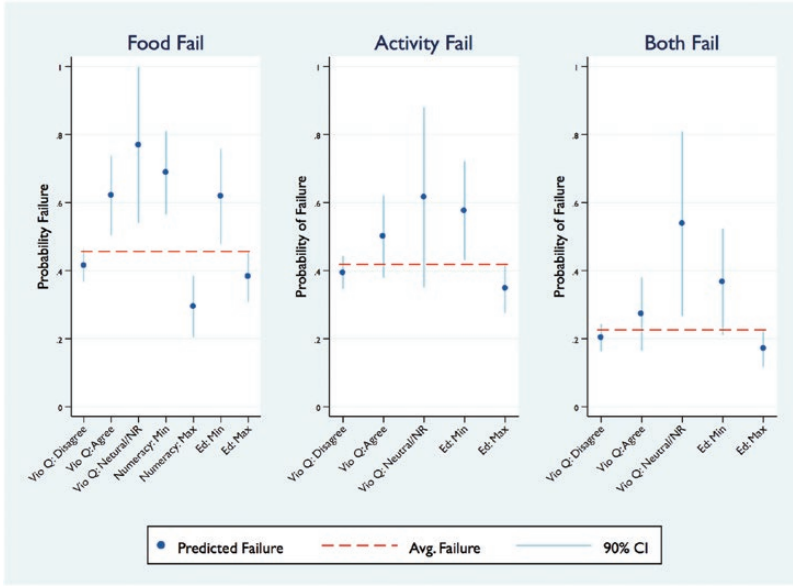

Figure 6. Nonsensitive list experiment failures.

Which Respondents Contribute Most to List Experiment Failure?

List experiment failures of the kind we have identified can be problematic in different ways. Failure is least damaging if distributed evenly across respondents. If LE response errors are systematically above *and* below the true value, the primary consequence is statistical power. This matters given the comparably lower statistical efficiency of the list experiment, but aggregate LE estimates will not be statistically biased. More problematically, bias may be systematic if violations are concentrated among certain population subgroups. If this group is one where the sensitive attitude/behavior is most prevalent, the LE may simply not work or we may come to biased conclusions about correlates of the sensitive attitude/behavior.

It is therefore important to understand which subgroups are more likely to incorrectly complete the list experiment. To do so, we estimate logistic regression models where the dependent variables are the dichotomous measures of LE failure. We use three outcomes: failure in the food list, failure in the activity list, and failure in both lists. We estimate the effect of education and numeracy separately without other covariates.[10]

Figure 6 summarizes the results, showing the predicted failure rates at minimum and maximum levels of our ordinal measures of education and numeracy. Individuals with less education and who engage in fewer numeracy activities are substantially more likely to provide inconsistent answers to the list and

10. Because our goal is to understand list experiment breakdown, we focus on subgroups where breakdowns are more likely, not ones substantively relevant to understanding our sensitive topic.

direct questions.[11] These patterns support the claim that the complexity and difficulty of the list experiment can produce measurement error.

We also examine how responses to the direct question about support for political violence correlate with performance on these simple lists. Figure 6 shows that respondents who agreed with the direct statement regarding support for election violence are significantly more likely to incorrectly complete the nonsensitive list experiment. Most critically, those who performed worst on the LE did not provide a response to the direction question regarding violence. The literature on sensitive surveys has long held that item non-response is a central sign of social desirability bias. In this application, the nonsensitive LE produces inconsistent answers among the population where the LE should be most useful.

DOES THE SENSITIVE TOPIC LIST EXPERIMENT WORK?

We now turn to the LEs that measure support for political violence. While food and activity list experiment "failures" were unambiguously attributable to error, respondents may intentionally misreport answers to sensitive LEs. Thus, for the political violence LEs, we characterize instances where the list experiment does not reveal prevalence estimates greater than asking outright as "LE breakdowns"—driven by list experiment failure and potentially also by deliberate response bias. Before proceeding to the main results, we evaluate how the conventional LE compares to the direct question.

Panel A of table 5 shows the distribution of numerical responses in the standard double list experiment (DLE). For both control lists, we observe that over 80 percent of respondents reply 2 or 3. Only one participant reported 0 and about 8 percent of respondents replied 4 (summing across Control 1 and Control 2 of table 5's Panel A), suggesting we generally avoided floor and ceiling effects.

Table 6 shows how the direct question compares to list experiment estimates derived from difference-in-means tests for the full sample and in relevant subgroups. In addition to education and numeracy considered in Section "Does the nonsensitive topic list experiment work?", we also include a measure of basic wealth and a proxy for political affiliation—ethnolinguistic group—where Luos and Kikuyus generally block vote in the opposition and governing party, respectively.[12] While our DLE design facilitates combining the two list experiments, we leave them disaggregated for more detailed analysis in this table.

In the table, we observe four patterns. First, consistent with existing literature (Droitcour et al. 1991; Ahart and Sackett 2004; Biemer and Brown 2005; Prior 2009; Tsuchiya, Hirai, and Ono 2007), we find that the standard list experiment procedure can produce a lower estimate of the sensitive attitude than direct questions (LE2; row 1). In this case, LE2 suggests that about 1 percent of

11. The education coefficient is significant a $p < .05$ for all three panels and numeracy at $p < .01$ in the first panel. Regression results are found in Online Appendix B.

12. While educational attainment of Kikuyus is higher on average nationally, politically relevant ethnicity is not correlated with education or numeracy in our sample (likely due to our focus in urban areas) and thus does not confound the relationship between education/numeracy and LE failure.

Table 5. The distribution of numerical responses in the sensitive list experiments

Panel A: Standard LE	Control 1	Treatment 1	Control 2	Treatment 2
0	0%	1%	0.4%	2%
1	9%	7%	8%	7%
2	40%	32%	42%	44%
3	41%	40%	42%	36%
4	10%	18%	7%	10%
5	–	1%	–	1%
Mean	2.52	2.69	2.47	2.48
N	234	242	240	235
Panel B: Tick LE	Control 1	Treatment 1	Control 2	Treatment 2
0	0%	0%	2%	1%
1	9%	5%	8%	9%
2	31%	28%	41%	36%
3	47%	44%	46%	43%
4	12%	20%	4%	10%
5	–	3%	–	2%
Mean	2.62	2.88	2.42	2.57
N	121	103	103	121
Panel C: Cartoon LE	Control 1	Treatment 1	Control 2	Treatment 2
0	0%	0%	1%	1%
1	9%	6%	7%	11%
2	39%	24%	46%	39%
3	35%	47%	42%	39%
4	18%	19%	5%	11%
5	–	4%	–	0%
Mean	2.61	2.91	2.43	2.46
N	114	139	137	114

NOTE.—Table presents the distribution of numerical responses in each list experiment. Numbers in parentheses indicate the percentage of respondents in each list experiment that gave each response. All subjects participated in the standard list experiment. Roughly half were randomly assigned to either the tick or the cartoon modifications.

respondents agree with the sensitive statement, compared to about 14 percent with the direct question. Sensitivity bias should push estimates from direct questions downward, so while the list experiment estimate is not significant, this is evidence of potential LE breakdown. We also see evidence of this in respondent subgroups. Second, we observe a "successful" list experiment within education subgroups (LE1; rows 3 and 5). Here, the technique suggests that 26–28 percent of respondents hold the sensitive attitude and the difference in mean applicable

Table 6. Conventional list experiment estimates

Variable	Value	Direct	LE1	LE2
Overall		0.14	0.17	0.01
			(0.08)	(0.07)
Education	Less than primary	0.25	–0.24	0.84#
			(0.60)	(0.43)
	Primary	0.17	0.28#	0.02
			(0.15)	(0.15)
	Some secondary	0.22	0.23	–0.14
			(0.21)	(0.20)
	Secondary	0.10	0.26*	–0.06
			(0.13)	(0.12)
	Tertiary	0.08	–0.14	0.20#
			(0.18)	(0.15)
Numeracy (Mean = 5.2)	0–4 Activities	0.16	0.13	–0.30*
			(0.13)	(0.12)
	5–9 Activities	0.13	0.19#	0.17#
			(0.10)	(0.09)
Income shock (Last Year)	Never	0.09	0.13	–0.11
			(0.13)	(0.12)
	Once	0.19	0.23	0.02
			(0.19)	(0.20)
	Many times	0.15	0.20#	0.07
			(0.11)	(0.11)
Ethnicity	Luo	0.31	0.08	0.16
			(0.20)	(0.18)
	Kikuyu	0.11	0.09	0.05
			(0.13)	(0.13)
	Other	0.11	0.26*	–0.07
			(0.12)	(0.11)

Standard errors of the list experimental estimates in parentheses.

# $p < 0.10$, * $p < 0.05$ for two-tailed Welch's T-test with unequal variance comparing treatment and control lists for each of the two list experiments.

items between treatment and control lists is significant. This pattern is observed for other respondent subgroups as well. Third, there are "smoking gun" breakdowns. For the low numeracy subsample, for example, LE2 suggests a prevalence of the sensitive attitude that is statistically significant and less than zero. Fourth, we are reminded of the efficiency costs of LEs. Some subgroups reveal greater prevalence of the sensitive attitude through the LE versus direct question, but the list experiment estimate is insignificant due to insufficient statistical power. While DLEs potentially reduce the efficiency cost, the table shows that

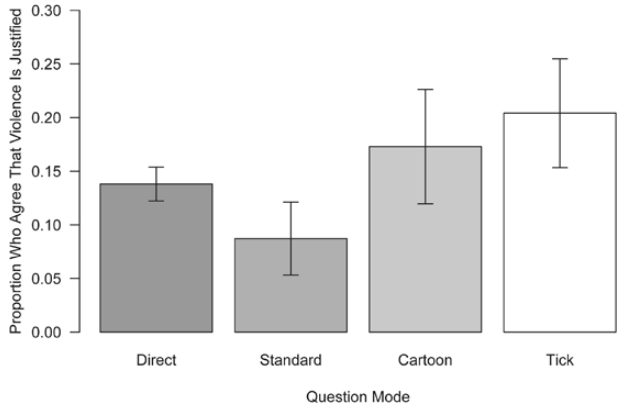

Figure 7. Full sample estimates of agreement with the sensitive violence statement. List experiments are all implemented using the DLE design (Glynn 2013). The bars present standard errors, calculated using the variance formula for the double list design (Droitcour et al. 1991).

combining divergent LE1 and LE2 estimates could actually introduce additional statistical noise.

DO LIST EXPERIMENT MODIFICATIONS IMPROVE PERFORMANCE?

We now introduce the results of the modified procedures. Panels B and C of table 5 present the distribution of numerical responses to each list in the modified list experiments. The high concentration of responses at 2 and 3 again suggests we avoided floor and ceiling effects.

Figure 7 displays a series of estimates of the proportion of respondents who agree that violence is justified if another ethnic group steals an election. The far-left bar represents our estimate from the direct question. About 14 percent of our respondents reported that they do agree with the statement when asked directly. The remaining bars show LE estimates. Because our goal is to assess the performance of our LE modifications relative to the standard list experiment, we show the standard procedure and each modification separately. Each is derived from DLE estimates in the manner discussed previously.

The second bar reveals that the standard LE produces an estimate that is lower than the direct question, evidence of potential measurement error in the standard procedure. By contrast, the final two bars demonstrate that the list experiments implemented using our modifications both produce estimates that are simultaneously higher than the direct item and which are statistically different from the standard procedure estimate. Using the cartoon procedure, we estimate that 17 percent of the sample agrees with the statement, while 20

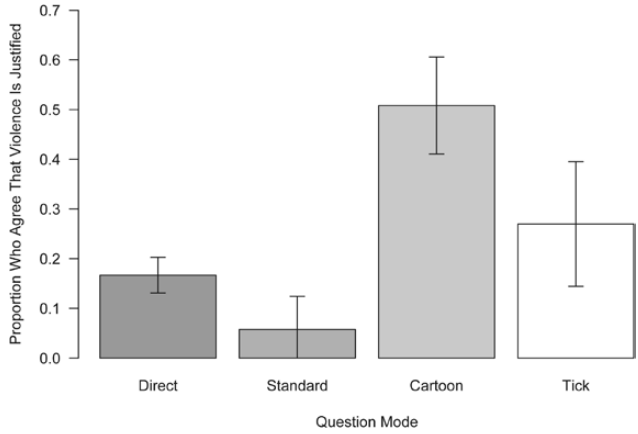

Figure 8. Estimates of agreement with the sensitive violence statement among nonsensitive list experiment failures. This chart presents the estimates by question mode for the 108 participants who did not match on both the food and activities lists.

percent of the sample agrees when using the tabulation procedure. Though we cannot statistically distinguish the modification results from the direct question, the modification estimates are statistically different from the standard procedure estimate, even though the sample of participants used to generate these estimates are identical.[13] We take this as evidence that our modifications improved the efficacy of the list experiment for our full respondent sample.

How do the innovations perform for those respondents who are most likely to have difficulty with list experiment questions? We address this question by shifting focus to the 108 individuals in our sample who failed to match on the food *and* the activities lists discussed in Section "Does the nonsensitive topic list experiment work?".[14] Figure 8 presents results from the list experiment modifications from this subsample. The figure illustrates that about 16 percent of those that failed both food and activity lists agreed with the sensitive statement regarding political violence when asked directly. In contrast, the standard LE procedure estimates a 5 percent prevalence of that attitude. This is well below that of the direct question and, notably, its confidence interval includes zero. Among these individuals within our sample, the conventional LE performs especially poorly.

13. Standard LE estimates are about the same in the subgroups that receive the cartoon and tabulation modifications. Additionally, the results are comparable, though statistically less efficient, when we restrict the sample to include only data from the first LE in which each respondent first participated.

14. In Online Appendix F.2, we present the distribution of numerical responses to the LE questions among this subgroup.

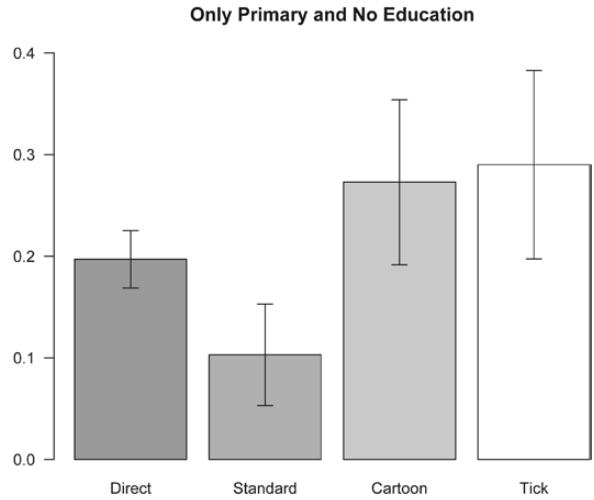

Figure 9. Estimates of agreement with the sensitive violence statement among those with only primary education or no formal education. This chart presents the estimates by question mode for the 132 participants who never attended school or whose highest level of attendance was primary school.

Our modifications, on the other hand, work well with this challenging group. The estimate increases substantially to 50 percent with the cartoons and 27 percent with tabulation. As there are only 108 "likely LE failures," our estimates with our modifications are very imprecise even with the DLE design. Even so, it is clear that they are substantially larger than the direct question, and much more reasonable than the standard procedure estimate, which again under-predicts the sensitive attitude in this subsample of respondents.

The results are comparable when we restrict the sample to include only those who either never attended school or whose highest level of education is primary school.[15] Figure 9 shows that the standard LE procedure underestimates support for violence versus the direct question. Our modifications, on the other hand, yield estimates higher than the direct estimate—exactly what one would expect from a list experiment given the sensitivity of the item of interest.

Discussion

This study makes several contributions. First, we show that list experiments can "fail" because of additional complexity and difficulty. At present, the burden of proof by which an LE is deemed "successful" is minimal. The convention in the

15. In Online Appendix F.3, we present the distribution of numerical responses to the LE questions among this subgroup.

literature is to "simply assume that if...[it] yields a greater prevalence of the sensitive behavior than asking directly, this is due to a reduction in response bias" (Simpser 2017, 2). Publication biases make reporting on failed list experiments rare, meaning we do not understand the broader distribution of list experiment "success." We contribute by providing evidence of "failure" and insight on the populations in which such failures are likely to be most concentrated.

Second, we offer practical lessons for survey researchers implementing list experiments. LE breakdowns are documented in varied settings. We introduced two cost-efficient LE modifications that reduce complexity and difficulty. We are optimistic that the modifications can work well in other research contexts and hope future research will test their potentially broad application. Our cartoon modifications are designed with less literate populations in mind. NGOs and aid agencies regularly use cartoons and visual aids to communicate complex information about elections, civic participation, and health behaviors—topics subject to sensitivity biases—throughout the developing world. Private tabulation could be used in settings where literacy rates are high. Our investigation of the modifications reinforces two related lessons for surveys in developing-country settings: (1) minimizing required effort is critical to data quality; and (2) concrete implementation details are essential to assess what techniques work best (Lupu and Michelitch 2018).

We conclude by reflecting on the study's implications for efforts to reduce survey error. The challenge of sensitive topics and behaviors has long beguiled survey research because respondents have strong incentives to self-censor. Given the potential costs of divulging illegal actions or undesirable attitudes, the risks simply are not worth being truthful. When such questions yield misrepresentation or item non-response, they introduce bias. The list experiment is an important tool that can help researchers collect more accurate survey data about sensitive attitudes and behaviors. While there is a growing apparatus of techniques for more efficiently designing and analyzing list experimental data, our paper has identified first-order questions about whether list experiments work and whether they may introduce new forms of error. Unlike predictable downward biases present with direct questions about sensitive attitudes and behaviors, list experiments may yield less predictable response error. Thus, in designing list experiments, survey researchers must consider potential tradeoffs between reducing the costs associated with honesty and increasing the costs associated with optimally answering.

Supplementary Data

Supplementary data are freely available at *Public Opinion Quarterly* online.

References

Adida, Claire, Karen E. Ferree, Daniel N. Posner, and Amanda Lea Robinson. 2016. "Who's Asking? Interviewer Coethnicity Effects in African Survey Data." *Comparative Political Studies* 49:1630–60.

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

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

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, Winston Chou, and Kosuke Imai. Forthcoming. "List Experiments with Measurement Error." *Political Analysis* 1–37.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2018. "When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments." Working Paper. Available at https://graemeblair.com/papers/sensitivity-bias/, accessed January 14, 2019.

Blair, Graeme, C. Christine Fair, Neil Malhotra, and Jacob N. Shapiro. 2013. "Poverty and Support for Militant Politics: Evidence from Pakistan." *American Journal of Political Science* 57:30–48.

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

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. "Design and Analysis of the Randomized Response Technique." Journal of the American Statistical Association 110:1304–19.

Carlson, Elizabeth. 2014. "Social Desirability Bias and Ethnic Voting on African Surveys." *Afrobarometer Working Paper Series* 144:1–28.

Central Bank of Kenya, Kenya National Bureau of Statistics, and FSD Kenya. 2016. "The 2016 FinAccess Household Survey." Available at http://fsdkenya.org/publication/finaccess2016/, accessed February 20, 2019.

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

———. 2018. "Clientelism in Competitive and Uncompetitve Elections." *Comparative Political Studies* 51:76–104.

De Cao, Elisabetta, and Clemens Lutz. 2018. "Sensitive Survey Questions: Measuring Attitudes Regarding Female Genital Cutting Through a List Experiment." *Oxford Bulletin of Economics and Statistics* 80:871–92.

de Jonge, Chad P. Kiewet, and David W. Nickerson. 2014. "Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys." *Political Behavior* 36:659–82.

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 Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 185–210. Hoboken, NJ: Wiley & Sons.

EU-EOM. 2007. "Final Report: General Elections 27 December 2007." Available at http://eeas.europa.eu/archives/eueom/pdf/missions/final_report_kenya_2007.pdf.

Finkel, Steven E., Jeremy Horowitz, and Reynaldo T. Rojo-Mendoza. 2012. "Civic Education and Democratic Backsliding in the Wake of Kenya's Post-2007 Election Violence." *Journal of Politics* 74:52–65.

Flavin, Patrick, and Michael Keane. 2009. "How Angry Am I? Let Me Count the Ways: Question Format Bias in List Experiments." Unpublished manuscript. Available at http://www.academia.edu/download/44601645/How_Angry_Am_I_Let_Me_Count_the_Ways_Que20160410-13091-1obdje.pdf, accessed January 14, 2019.

Frye, Timothy, Scott Gehlbach, Kyle L. Marquardt, and Ora John Reuter. 2017. "Is Putin's Popularity Real?" *Post-Soviet Affairs* 33:1–15.

Gelman, Andrew. 2014. "Thinking of Doing a List Experiment? Here's a List of Reasons Why You Should Think Again." Available at http://andrewgelman.com/2014/04/23/thinking-list-experiment-heres-list-reasons-think, accessed February 20, 2019.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77:159–72.

Gonzalez-Ocantos, Ezequiel, Chad Kiewet de Jonge, Carlos Melendez, Javier Osorio, and David W. Nickerson. 2011. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56:202–17.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74:37–67.

Humphreys, Macartan, and Jeremy Weinstein. 2006. "Handling and Manhandling Citizens in Civil War." *American Political Science Review* 100:429–47.

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.

Kasara, Kimuli. 2013. "Separate and Suspicious: Local Social and Political Context and Ethnic Tolerance in Kenya." *Journal of Politics* 75:921–36.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Journal of Cognitive Psychology* 5:213–36.

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

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.

Lax, Jeffery R., Justin H. Phillips, and Alissa F. Stollwerk. 2016. "Are Survey Respondents Lying About Their Support for Same-Sex Marriage? Lessons from a List Experiment." *Public Opinion Quarterly* 80:510–33.

Lupu, Noam, and Kristin Michelitch. 2018. "Advances in Survey Methods for the Developing World." *Annual Review of Political Science* 21:195–214.

Lyall, Jason, Graeme Blair, and Kosuke Imai. 2013. "Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107:679–705.

Oliveros, Virginia. 2016. "Making it Personal: Clientelism, Favors, and the Personalization of Public Administration in Argentina." *Comparative Politics* 48:373–91.

Prior, Markus. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *Journal of Politics* 71:893–908.

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

Scacco, Alexandra. 2016. "Anatomy of a Riot: Participation in Ethnic Violence in Nigeria." Book manuscript, New York University. Available at: http://www.nyu.edu/projects/scacco/files/Scacco_Anatomy_of_a_Riot.pdf, accessed February 20, 2019.

Simpser, Alberto. 2017. "When do Sensitive Survey Questions Elicit Truthful Answers? Theory and Evidence with Application to the RRT and the List Experiment." Available at SSRN: https://ssrn.com/abstract=3032684 or http://dx.doi.org/10.2139/ssrn.3032684

Singer, Elanor, Hans-Jurgen Hippler, and Norbert Schwarz. 1992. "Confidentiality Assurances in Surveys: Reassurance or Threat?" *International Journal of Public Opinion Research* 4:256–68.

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

Suri, Tavneet, and William Jack. 2016. "The Long-Run Poverty and Gender Impacts of Mobile Money." *Science* 354:1288–92.

Tourangeau, Robert. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by Thomas B. Jabine, Miron L. Straf, Judith M. Tanur, and Robert Tourangeau, 73–199. Washington, DC: National Academies Press.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–93.

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

Zigerelli, Lawrence J. 2011. "You Wouldn't Like Me When I'm Angry: List Experiment Misreporting." *Social Science Quarterly* 92:552–62.