(Mis)Measuring Support for Election Violence with the List Experiment:
Evidence from Kenya

Eric Kramon[*]
Keith Weghorst[†]

March 15, 2018

Word count: 6,498

[*] Department of Political Science, George Washington University. Email: ekramon@gwu.edu
[†] Department of Political Science, Vanderbilt University. Email: keith.weghorst@vanderbilt.edu

Abstract

List experiments 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 question 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, non-sensitive lists about food consumption and daily activities: over 40% 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 they 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 "failure." We then 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.

1 Introduction

Survey researchers are often concerned with measuring sensitive attitudes and behaviors, including support for political violence, experience with corruption, or racial attitudes. A major challenge in studying such topics with direct survey questions is their sensitivity; individuals often do not want to reveal objectionable actions or attitudes because of social desirability bias or fear of legal sanction. To circumvent this challenge, survey researchers have developed a number of strategies for reducing measurement error driven by sensitivity biases. The list experiment — also called the "item count technique" — is one such strategy that is increasingly popular in political science and related disciplines due to its broad applicability.

List experiments reduce survey error by asking respondents about sensitive issues indirectly: by embedding sensitive items among multiple non-sensitive items and asking respondents to aggregate the total number of applicable items (but not which items). By less invasively asking about sensitive topics, the technique reduces the perceived costs/risks of answering honestly. However, the enthusiasm surrounding the list experiment has drawn attention from its potential limitations. In particular, the length and complexity of the question format make them costlier for respondents to complete "optimally,"[1] and thus they are prone to comprehension and reporting errors. More critically, such errors may be systematically concentrated among certain population subgroups, including those without experience answering complex survey questions, or even those in which the sensitive behavior/attitude of interest is most prevalent. 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 list experiments often "fail" in ways that are difficult to miss: for example by producing estimates of the frequency of a sensitive attitude that are lower than rates provided by direct question, or even results that are non-sensical, such as those that are negative (Holbrook and Krosnick, 2010). *Why*

[1] See Krosnick (1991).

do list experiments sometimes "fail"? How can the performance of the list experiment be improved so that it is useful for survey researchers and pollsters?

In this paper, we examine the list experiment and its ability to reduce survey error in Kenya, where we implemented a study designed to measure public support for political violence. First, we investigate the performance of the list experiment using lists of simple, non-sensitive items about food consumption and daily activities. We show that the list experiment encounters difficulties with these simple and non-sensitive lists: over 40 percent of respondents provide inconsistent responses in the list and direct question formats. These list experiment "failures" are concentrated among less numerate and less educated respondents, evidence that these errors are driven by list experiment question complexity and difficulty. They are also most common among participants who refuse to answer the direct question about political violence, diluting the core benefit of using the list experiment.

The second empirical section of the paper turns to list experiments designed to measure attitudes about political violence. We find that the standard list experiment reveals lower rates of support for political violence than those obtained by simply asking directly about this sensitive attitude. These under-estimates are most pronounced among less educated participants and those respondents who provided inconsistent responses in the non-sensitive list experiments described above. To test potential solutions to the problem, we evaluate two modifications to the list experiment that are designed to reduce the complexity of the technique (and which are appropriate in the Kenyan context). The first allows for private tabulation, and combines private tabulation with cartoon visual aids. We find that both modifications improve the performance of the technique.

The paper contributes to the literature on survey response bias and the measurement of sensitive political attitudes and behaviors in two central 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. Our work is among the first to evaluate the list experiment and directly identify its breakdown, providing evidence of what drives list experiment failure.[2] The survey literature is populated with list experiments that perform well; we highlight limitations that might not be obvious from reading the published literature due to publication bias. Our aim is not to suggest that all list experiments are problematic, but rather to draw attention to these limitations.

Our second contribution is demonstrating 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. We conclude the paper by identifying future opportunities to further study and enhance list experiment performance, and by discussing the implications of our results for survey researchers.

2 Measuring Sensitive Attitudes with the List Experiment

Attitudes toward violence are both commonly studied in the social sciences and emblematic of the challenges of studying sensitive topics. Like a number of important topics, attitudes towards political violence are subject to under-reporting biases because such violence is illegal and supporting it is generally socially undesirable. Past research on violence has addressed sensitivity-driven measurement error by enhancing the secrecy of responses, thereby reducing perceived costs/risks of answering questions truthfully. Strategies include not asking about violent behavior directly (Humphreys and Weinstein, 2006), administering sensitive survey modules separately from a larger survey (Scacco, 2016), anticipating or controlling for the identity dyad of respondents and enumerators (Adida et al., 2016; Carlson, 2014; Kasara, 2013), or one of several experimental approaches: endorsement experiments (Blair et al., 2014; Lyall, Blair and Imai, 2013), randomized response technique (Blair, Imai and Zhou, 2015), or the list-experiment/item count technique.

[2] To our knowledge, there exists to date only one empirical validation study of the list experiment. It compares voting behavior to turnout records and shows that the technique performs worse compared to alternative indirect measures (Rosenfeld, Imai and Shapiro, 2016).

The list experiment, also called item-count technique, is a promising alternative to direct questions, offering interviewees greater secrecy for sensitive responses (e.g. Blair and Imai, 2012; Corstange, 2010; Glynn, 2013; Gonzalez-Ocantos et al., 2011; Kuklinski, Cobb and Gilens, 1997). The list experiment presents a sensitive statement as one of many items of a list and asks respondents to identify how many total items from the list apply to them. Participants are randomly assigned to either a treatment list that includes 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 responses to the 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.

The literature emphasizes two assumptions that must be satisfied for list experiment estimates to be valid: the "no-liars" and "no design effect" assumptions (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 that respondents engage control items. List are generally designed to avoid "floor" and "ceiling" effects, which undermine the goal of keeping individual item responses undetectable (Glynn, 2013).[3]

For single list experiments, 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 a double list experiment design, the estimate is a weighted average of the two single list experiment estimates that compose it (Glynn, 2013).

2.1 Do List Experiments Work?

How effective are list experiments? Demonstrating that a list experiment "works" is difficult because the true prevalence of a sensitive item of interest is unknown. Discussing the list experi-

[3] In addition, some randomize the position of the sensitive item on the treatment lists in order to avoid design effects related to the position of the sensitive item.

ment, Simpser (2017) notes that “it is not currently known whether, or under what circumstances, question structures designed to elicit sensitive information surveys actually work” (3). To our knowledge, there exists only one study of list experiment performance with a sensitive item where objective validation is possible; it finds the list experiment reduces response bias but with greater efficiency costs than direct questions and other indirect alternatives (Rosenfeld, Imai and Shapiro, 2016). When comparisons to direct questions are possible, the list experiment has a concerning record of mixed success.

First, because of their inefficiency, successful list experiments may not be worth the magnitude of sensitivity bias they manage to offset (Tourangeau and Yan, 2007). Second, list experiments may reveal responses that are statistically similar or not differentiable from direct questions Third, list experiments even produce under-estimates of sensitive behaviors and attitudes.

One overview of 48 studies with list experiments, finds that only 63% of them “worked.” The list experiments in the remaining studies “failed” in that they either were either statistically indistinguishable or revealed *lower prevalence* of the sensitive attitude compared to direct questions (Holbrook and Krosnick, 2010). Such failures are observed across sensitive topics, including drug and alcohol use (Biemer and Brown, 2005; Droitcour et al., 1991; LaBrie and Earleywine, 2000), shop-lifting (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 employee theft (Ahart and Sackett, 2004). Some results are non-sensical —in one study, a list experiment estimated 2.4% of respondents reported having been “abducted by extraterrestrials” (Ahlquist, Mayer and Jackman, 2014). List experiment failures may predominate in certain population subgroups (Zigerelli, 2011), which can be especially inferentially damaging under some circumstances.

These findings raise important concerns about the list experiment in various settings. A number of scholars have also raised apprehension over a “file-drawer” problem with list experiments whereby most list experiment failures are never reported.[4]

2.2 Why Do List Experiments Break Down?

What explains the mixed record of success for the list experiment? Most research has focused on how the technique reduces the potential costs of respondent honesty regarding a sensitive item. Our paper focuses on a fundamental issue with the list experiment: that the process of answering list experimental questions is costlier to respondents in that it is more complex and difficult than answering direct questions. We argue this can introduce unanticipated measurement error that, unlike downward sensitivity biases, operates in a less predictable manner—including both upward and downward biases (de Jonge and Nickerson, 2014). Indeed, answering survey questions is demanding. To complete a question "optimally," respondents must interpret the meaning of a question, comb their memories for all information relevant to the question, aggregate the information into general views, and report these general views with precision (often having to translate views into presented response options) (Krosnick, 1991).[5] Respondents may put forth different levels of effort to engage, follow instructions, and correctly answer list experiments versus direct questions. If so, list experiments might introduce a cost distinct from direct questions which may introduce distinctive measurement error. The core issue underlying list experiment failures is this "honesty-effort compromise," whereby list experiments reduce perceived costs of being honest but require significant additional effort to actually answer truthfully.

List experiments are complex and more costly to optimally answer than conventional questions for several reasons. First, instructions for the list experiment are more extensive and less familiar than most question prompts. Even understanding the list experiment procedure requires more effort than conventional questions. Second, list experiments often use control items that probe about experiences (health activities, past social interactions, etc.) or attitudes requiring

[4] See, for example, this discussion on Andrew Gelman's blog.

[5] Paraphrasing Krosnick (1991)'s discussion of Tourangeau (1984)'s definition of interviewees *optimizing* survey responses.

longer-term recall. Third, respondents must perform the higher-order task of individually considering a set of qualitative statements of varying complexity and the arithmetic task of adding up items on a list.

We argue that the difficulty of these features can lead to list experiment failure, particularly in low-development settings where the technique is increasingly implemented. Such failures are likely to be concentrated in groups of individuals with less education and less skill to perform the tasks required by the list experiment. The question format may be the most unfamiliar and challenging for these individuals and thus the costs of effort required to complete the list experiment may undermine the reduced costs/risks the list experiment provides to facilitate honest reports of sensitive items. Scholars are correctly concerned with the "bias-variance trade-off" that exists when choosing direct questions versus indirect formats like the list experiment (Rosenfeld, Imai and Shapiro, 2016). Our concern is more fundamental: achieving honesty with the list experiment requires significant additional effort and that effort may be systematically more demanding for some respondents. Thus, applied researchers using the list experiment may in fact face a "bias-bias trade-off."

Evidence from a failed list experiment carried out in Tanzania is suggestive of this intuition. In this setting, we introduced a list experiment to measure attitudes toward the statement: "Under certain circumstances, the political opposition must use violence against the government."[6] Using a measure of educational attainment, Figure 1 compares direct question estimates to list experiment estimates across three educational attainment categories.[7] The list experiment performs especially poorly for the respondents with the least amount of formal education. Notably, for those in low and middle education groups, the list-experiment estimate is very close to 0 or even negative, and well below the estimates generated by the direct questions. As we would expect downward bias on direct questions about support for political violence, these dramatic underestimates of the list

[6] Additional information on question wording is found in Appendix 2.

[7] We define "low education" as those with no formal schooling or only primary schooling, "middle education" as those with only secondary schooling, and "high education" as those with tertiary or university education.

Figure 1: **Support for Oppositon Violence in Tanzania.** Estimates from the List Experiment and Direct Question.

FIGURE HERE

experiment suggest a significant problem.

For the study of election violence and political violence generally, results such as these raise alarms about the list experiment. This is one topic where respondents who hold the sensitive item of interest are concentrated in the very population where the "honesty-effort compromise"—that is, where the effort required to report honestly through the list experiment—is greatest. Many topics studied with the list experiment are also likely to be most prominent among lower education, literacy, or socioeconomic status levels. Given these potential challenges, we present a study designed to detect such failures and modifications to the list experiment that can improve its performance by reducing the effort required to accurately complete list experiments.

3 Research Design

We conducted this research in Kenya ahead of the 2013 elections. Attitudes about ethnic violence are an especially sensitive topic in Kenya given country's experience with large-scale post-election violence. While often portrayed as driven by interethnic animosities, roots of conflict lie in the politicization of the identities through political competition and resource inequalities (Kagwanja, 2003; Kanyinga and Long, 2012; de Smedt, 2009). In what follows, we describe our strategy for assessing the performance of the list experiment in Kenya, and two modifications proposed to enhance the list experiment's performance in low-development settings.

3.1 Survey Design

In June 2012, we tested the performance of the list experiment and possible modifications in Kenya on the topic of election violence. We conducted the survey in four areas within the capital of Nairobi — Githurai, Karangware, Kibera, and Mathare — that experienced violence and unrest in 2007-08.[8] Our goal was to measure beliefs regarding the following sensitive statement: *If another tribe tries to steal an election, it is justified to use violence to try to stop them.*[9] We chose this wording deliberately; most people object to violence in principle. Our goal was to measure whether people believe that intergroup violence is justified in electoral contexts, not the extent to which they believe it is "good." The language also tapped into perceptions that election malfeasance plays a large role in political violence in Kenya.

3.1.1 Detecting Failures due to Survey Item Complexity and Difficulty

Our initial aim is to assess whether the 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 list experiment formats. Violations should be observable when responses to direct questions about list items do not match the numerical response provided when the items are posed in list format. For example, if a respondent responds "3" to the list-experiment question, they should agree with three of the statements when asked directly. However, because respondents have strong incentives to misreport about sensitive topics, assessing the performance of list experiments that include sensitive items is difficult. In standard list experiments, the potential response bias on the

[8] More detail on our sample design and procedure can be found in Appendix 2.

[9] We understand that the term "tribe" has negative connotations in many settings. In Kenya, however, people speak of "tribalism" and "tribes," not ethnic groups and ethnic identities. Working with our research team, we translated the statement into Sheng, a swahili-based slang dialect that includes elements of English and some other Kenyan languages. Sheng is widely spoken in Nairobi, and is more easily understood than formal Swahili, which most Kenyans learn as children but rarely use in daily life.

direct sensitive item makes it difficult to interpret mismatches between numerical responses and direct item responses. In addition, when list experiment items are relatively complex or require substantial thought, responses to list and direct items may differ as an artifact of the technique itself (e.g. Flavin and Keane, 2009; de Jonge and Nickerson, 2014). Detecting list experiment failures is therefore challenging with existing list experiment data, as they almost universally address topics subject to sensitivity biases. Lacking some other, non-survey measure by which to validate list experiment respondents (See: Rosenfeld, Imai and Shapiro, 2016), we must use non-sensitive topics which are not subject to such response biases.

We designed an experiment with the explicit goal of assessing list experiment failure. Respondents were prompted with instructions common to most list experiments and then read lists of very simple activities from daily life.[10] The items were entirely non-sensitive so that direct questions about them would not be subject to response bias. They were also simple in order to make recall easy and to ensure that direct item responses would not differ from list responses, as in Flavin and Keane (2009). The first and most basic non-sensitive list asked participants about foods consumed in the prior week. All of these foods are commonly consumed in Kenya, with varying frequency. In the second non-sensitive list, we presented respondents with a number of everyday activities which they may engage in. Some of these activities capture counting and enumerating, like sending currency via mobile phone ("Mpesa") or depositing money into a bank. Table 1 presents the wording used to asked about food consumption and activities and the non-sensitive items included on the two lists.

TABLE 1 HERE

Less than five minutes after reading the lists, enumerators administered direct questions about food consumption and everyday activities. To make the food questions as simple as possible, respondents viewed pictures of fourteen total food items with their Swahili name underneath,

[10] Specific wording is found an Appendix 2.

including the five included on the list experiment. Interviewees used tablets to tap each food they had eaten in the previous week and the tablet recorded which items they selected. Direct questions corresponding with activities were designed to be more deliberate, so we did not provide respondents with a visual aide, and instead asked about each activity orally (one-by-one). Participants were asked whether they had engaged in each of the five activities, as well as four additional ones that require numeracy in order to generate a measure of numeracy skill and experience. In addition to detecting list experiment failures, we use these activities to construct a numeracy measure later in the paper.[11]

This procedure will allow us to demonstrate how realistic it is to expect respondents to put forth necessary effort for list experiments in Kenya. However, it does not provide solutions for alleviating list experiment failures. Thus, the study also analyzes the efficacy solutions for the list experiment when it does break down.

3.1.2 List Experiment Modifications

We now introduce two modifications to the list experiment designed to reduce the complexity and difficulty, developed with particular focus on the difficulties that low-education, low-numeracy respondents may experience. Our analysis of the modifications serves two purposes. First, if the list experiment performs better when the modifications are used, this would provide evidence that the list experiment is failing at least in part because of the difficulty respondents have with the question. Second, the modifications provide easy-to-implement solutions to the challenges we identity above.

3.1.2.1 List Handout and Private Tabulation In our first modification, we presented respondents with a textual aid and provideed them the opportunity to privately tabulate their answer. The

[11] From our previous surveys, we found that conducting numeracy skills tests such as asking respondents to answer simple and complex math problems sometimes offended respondents. Further, arithmetic problems only address one specific component of the list experiment procedure (aggregation) and we believe comfort with translating information from words to numbers is better captured by a wider range of numeracy related activities.

objective is to increase retention of each list item and to encourage deliberate tabulation. In the procedure, respondents were handed a laminated sheet of paper which included 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 understood this, the enumerator turned 90 degrees away from the respondent, read the list experiment instructions, and the lists. Upon completing the question, the enumerator returned 90 degrees to face the respondent and collected the material. Text prompts for the procedure are found in Appendix 2.

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 from respondents, our modifications also enhanced privacy. Thus, our modification design was a bundled treatment that also potentially reduced perceived costs of answering honestly, which we discuss in more detail later in the paper. Literacy is required, meaning some population subgroups will benefit less from this modification.

3.1.2.2 Cartoon Handout Our second innovation provides respondents cartoon visual aides. A local cartoonist created an illustration corresponding each list item. Each cartoon was placed on a handout and numbered by its ordering in the list experiment (e.g. the first item on the list was labeled "1"). After providing the respondent with the handout, the enumerator read the script for each list. 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 level to cognitive load with direct questions. This approach also further eased the aggregation process by allowing respondents to count the pictures corresponding with their attitudes,

Figure 2: Cartoon Corresponding with Sensitive Item, Kenya

FIGURE HERE

rather than reviewing written statements. Essentially, this modification replicates the private tabulation modification but with visual images rather than text, making it appropriate for participants with limited literacy skills. Figure 2 provides an illustration of the sensitive item. In Appendix 3, we present the cartoons that correspond to each list.

3.2 Survey Implementation

Our research design permits direct comparisons of the performance of our two modifications with the list experiment's standard implementation. Each respondent participated in the exact same list experiment at two different points in the survey using two different procedures: one of the two modifications and the standard procedure. We can therefore compare how individuals and sub-groups respond to different implementation procedures. The design also allows us to assess the performance of our two modifications to one another through between-group comparisons. The study enables us to examine how list experiment modifications perform for individuals who struggle the most with the conventional list experiment, based on failing the non-sensitive list experiment.

Of the survey's 478 participants, we randomly assigned 253 to the cartoon modification group and the remaining 225 to private tabulation. As each respondent participated in list experiments that were identical except in the manner that they were implemented (standard versus cartoon, or standard versus tabulation), we distracted participants after the first list experiment with unrelated questions, including the food and activities component. To control for ordering effects, we randomized the order in which the two list experiments are implemented.

Appendix 1 provides information about the sample and covariate balance. It shows that random assignment succeeded in distributing respondents evenly into each of the four potential list-experiment orderings and each double-list experiment sub-order. 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. [12] Appendix 1 shows that respondents in the randomly assigned groups are roughly comparable on a range of observable covariates, including gender, age, and education.

4 Results

We first determine whether the list experiment is likely to fail in settings like Kenya. We do so by testing list experiment performance through non-sensitive lists about food consumption and daily activities.

4.1 Results (1): Does the List Experiment Work?

Recall that we test for list experiment failures by comparing responses given to direct questions about non-sensitive activities to those same items in the list experiment format. Figure 3 shows whether respondents offered the same number of food and activity items when asked through the list experiment design versus direct questions. The x-axis shows the difference between the number of foods/activities reported through the list experiment design and the total reported from the individually posed direct questions. A value of 0 indicates responses for direct and list items

[12] To permit direct comparison between our modifications and the standard procedure, each individual receives the same treatment placement for both of the list experiments in which they participate. In other words, participants participated in list experiments that were exactly the same, except that the mode of implementation varied.

Figure 3: **Comparing estimates from the List Experiments and Direct Questions.**

FIGURE HERE

were the same. Positive values occur when the list response exceeded the number of items reported directly; negative values the opposite. Strikingly, less than 60 percent of respondents provide consistent information elicited from these two forms. The mismatch between the two question formats operates in both upward and downward directions, meaning that respondents both over- and under-predict behaviors when elicited through the list experiment design.

These findings are potentially troubling. Over 40 percent of our sample provides differing responses to list experiment and direct questions about innocuous aspects of their day-to-day lives. Even when queried about behavior as basic and non-sensitive as food consumption, and when provided with direct items in the most simple form possible — pictures that they look at and then touch — many respondents do not provide the same information. 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 respondents violate the basic efforts which must be met in order for respondents to honestly report answers for list experiment designs: that respondents consider each item in the way they do direct questions and then accurately aggregate the applicable statements.

We may have identified a lower bound on this problem. These list experiments were designed to be as simple as possible and were only separated from direct questions by a few minutes during the survey. Typical applications of the list experiment — probing attitudes about violence, fraud, vote-buying, racial prejudice, and so on — are far more complex.

4.1.1 Who does the list experiment break down for?

List experiment failures of the kind we have identified can be problematic in different ways. This failure is least problematic if such breakdowns are distributed evenly across study participants. If the number of items reported in response to a list experiment is just as likely to be higher than the true value as it is lower, the primary cost is statistical power. While this is important given the comparably lower statistical efficiency of the list experiment procedure, aggregate list experiment estimates of the sensitive item will not be statistically biased. More problematically, bias may be systematic in that violations may be concentrated among certain population subgroups where the list experiment is particularly challenging. If this group is one where the sensitive attitude/behavior are most prevalent, the list experiment may simply not work. Or, we may come to biased conclusions about the correlates of the the sensitive attitude/behavior.

For this reason, it is important to understand whether there are respondent types who are more likely to incorrectly complete the list experiment. First, are there attributes of respondents related to the difficulty of completing the list experiment that might lead to failures? Second, are list experiments likely to fail for individuals for whom direct questions are an inadequate measurement tool—those who refuse to answer such questions or otherwise lie when providing a response? If either is the case, then list experiments might either produce less efficient estimates of the sensitive item or increase survey error in a way that (or may not) introduce bias.

Figure 4 addresses the first question, showing the predicted rate of list experiment failures across education and numeracy levels. The values are generated based on logistic regressions for each of the three failure dependent variables — the food list, activity list, or both — and considering education and numeracy separately without other covariates. The figure demonstrates that individuals with less education and who engage in fewer numeracy activities in their daily life are substantially more likely to provide inconsistent answers to the list and direct questions.

Figure 4: **Non-Sensitive List Experiment Failures**

FIGURE HERE

Assuming that responses to the direct question are most accurate, these patterns are consistent with the claim that the complexity and difficulty of the list experiment can produce measurement error.

How do responses to the direct question about support for political violence correlate with performance on these simple lists? Figure 4 shows that respondents who agreed with the direct statement regarding support for election violence are significantly more likely to incorrectly complete the simple list experiment. More critically, those who performed the worst did not report an attitude at all. The literature on sensitive surveys has long held that item non-response and otherwise not meaningfully answering a question is a central sign of response bias. Thus, in this application, the simple list experiment is producing inconsistent answers among the population where the list experiment should be most useful.

4.2 Results (2): Support for Election Violence

We now turn to our examination of the list experiments that measure support for political violence. The research design is such that each participant responded to two sets of identical lists, one administered in the conventional mode and one with a modification. Before proceeding to the main results, we ask: how does the conventional list experiment perform compared to the direct question? Table 2 highlights these findings. While our double-list experiment design facilitates combining the two list experiments, we leave them disaggregated for more detailed analysis in this table.

TABLE 2 HERE

In the table, we observe three patterns. First, consistent with our previous findings (see Figure 1) and with others in the literature (e.g. Ahart and Sackett, 2004; Biemer and Brown, 2005; Droitcour et al., 1991; 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, the list experiment estimate suggests that about 1 percent of the population agrees with the sensitive statement, compared to about 14 percent with the direct question. Sensitivity bias should push estimates from direct questions downward, so when list experiments estimate sensitive topics at rates lower than direct questions, we believe these to be failures. We also see evidence of this in respondent subgroups.[13] Second, we observe a "successful" list experiment for the pooled respondent group (LE1; row 1). Here, the technique suggests 17% of respondents hold the sensitive attitude and the difference of mean applicable items between treatment and control lists one of the two list experiments is significant. This pattern is observed for some respondent subgroups as well. Third, there are "smoking gun" failures. For the low numeracy subsample, for example, the list experiment (LE2) suggests a prevalence of the sensitive attitude that is statistically signifiant and less than zero.

4.3 Results (3): List Experiment Modifications

We now integrate the results of the modified procedures. Figure 5 presents 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 presents our estimate from a direct question in which we ask respondents directly whether they agree or disagree with the statement. As we showed in Section 4.2, about 14 percent of our respondents reported that they do agree with the statement when asked directly. This proportion is higher than we had expected given how sensitive issues of violence and ethnic

[13] Many of the list experiment subgroup estimates are not statistically significant because we do not have sufficient power for subgroup analyses given the inefficiency of list experiment estimates.

politics are in Kenya. The second bar from the left presents the estimates produced by the double list experiment using the standard implementation procedure (pooling LE1 and LE2 from Table 2).

The first two bars show that the standard list experiment procedure produces an estimate that is lower than the direct item estimate, evidence of potential measurement error in the standard question. By contrast, the list experiments implemented using our modified procedures both produce estimates that are higher than the direct item and which are statistically different from the standard procedure estimate. The final two bars in Figure 5 present these estimates. Using the cartoon procedure, we find that about 17 percent of the sample agrees with the statement, while 20 percent of the sample agrees when using the tabulation procedure. Though we cannot distinguish the modification estimates from the direct question estimate, the modification estimates are statistically different from the standard procedure estimate, even though the sample of participants used to generate these estimates are identical.[14] We take this as evidence that our modifications improved the performance of the list experiment for our full respondent sample.

Figure 5: **Full Sample Estimates of Agreement with the Sensitive Violence Statement.** List experiments are all implemented using the double list experiment design (Glynn, 2013). Bars present standard errors, calculated using the variance formula for the double list design of Droitcour et al. (1991)

FIGURE HERE

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

[14] The standard list experiment estimates are about the same in the sub-groups 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 list experiment in which each respondent first participated.

Figure 6: **Estimates of Agreement with the Sensitive Violence Statement Among Non-Sensitive List Experiment Failures** This chart presents the estimates by question mode among the 108 participants who did not match on both the food and the activities list.

FIGURE HERE

Figure 7: **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 among the 132 participants who never attended school or whose highest level of attendance was primary school.

FIGURE HERE

our sample who failed to match on the food *and* the activities lists discussed in Section 4.1. Figure 6 presents results from the list experiment modifications from this subsample. The figure shows that about 16 percent of those that failed on both food and activity lists agree with the sensitive statement regarding political violence when they are asked directly. In contrast, the standard list experiment 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 list-experiment procedure performs especially poorly.

Our modifications, on the other hand, appear to perform 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 double list design. Even so, it is clear that they are substantially larger than the direct item estimate, 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 individuals who either never attended school or whose highest level of education is primary school. In previous sections, we provided evidence from a number of list experiments that showed systematic underestimation among this sub-sample of low educated respondents. Figure 7 again shows that the standard list experiment procedure under-estimates belief in the sensitive item relative to the direct question. Our modifications, on the other hand, produce estimates that are higher than the direct question estimate—exactly what one would expect from a list-experiment estimate given the sensitivity of the item of interest.

5 Discussion

Our paper has identified first-order questions about the performance of the list experiment. While list experiments are designed to reduce the perceived costs of reporting about sensitive attitudes honestly, they require additional effort in order to do so. Through a non-sensitive list experiment, we demonstrated that many respondents are either unable or unwilling to pay the additional costs of accurately reporting information in the list experiment format. Our paper has attempted to highlight how these additional costs can lead to new forms of measurement error, particularly among populations for whom the list experiment procedure is likely to be most unfamiliar and challenging.

In addition to showing how the complexity and difficulty of list experiments can lead to error, we also introduced two modifications — a cartoon handout and private tabulation — designed to address these challenges. We show that each of these modifications produces estimates of support for political violence that are more reasonable than the estimates produced by the standard procedure. Importantly, the modifications perform well when implemented with the sub-groups that have the most difficulty with the standard list experiment procedure. These results provide evidence for our claim that list experimental item complexity and difficulty can create measurement error, while identifying the effectiveness of easy-to-implement solutions to the problem.

Before discussing the implications of the findings, we note that one potential limitation of our research design is that, in implementing the modifications, enumerators turned away from respondents while they completed the list experiment and in doing so provided participants an additional level of privacy. Our modifications are thus a bundled treatment that simultaneously reduces cognitive complexity and enhances the privacy of reporting the list experiment. This could lead us to mistakenly attribute the success of our modifications to the reduction in required cognitive effort, when in fact the results are driven by the additional layer of privacy. We are confident that this risk is minimal because of the specific respondent subgroups for which the modifications improved list experiment performance.

Indeed if added privacy were responsible for the improved performance of the modified list experiments, then they would have addressed violations of the "no liars" assumption. Although we cannot know the true distribution of the sensitive attitude, we expect such improvements to be distributed fairly evenly across our respondents. By contrast, if our modifications help with cognitive effort, the improvements should be concentrated among certain respondents. To this end, we note three related pieces of evidence. First, our modifications performed especially well among respondents who had lower levels of educational attainment and numeracy. These two covariates were drivers of failure to match on the non-sensitive experiment regarding foods and daily activities. Second, our modifications performed best among those individuals who failed to match direct and list experiment responses for the non-sensitive list experiment. Modifications that enhance privacy should not be systematically associated with improvement among individuals who did not correctly complete the non-sensitive list experiment. Third, we note that some research has found that additional privacy provisions and reminders might actually introduce additional survey bias (Singer, Hippler and Schwarz, 1992). We leave future research to further investigate the impact that further enhancing the privacy of he list experiment has on the technique's ability to address "no liars" violations.

This potential limitation notwithstanding, this study makes several contributions. First, we show that list experiments can "fail" because of the additional complexity and difficulty of the question format. This is important because, at present, the burden of proof by which a list experiment is deemed "successful" is minimal. The convention in the 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).[15] Furthermore, publication bias make reporting on failed list experiments rare (Holbrook and Krosnick, 2010), meaning we have little understanding of the broader distribution of "success" of the list experiment as a survey technique. We contribute by providing evidence of "failure," and providing evidence about which types of populations such failures are likely to be most concentrated.

Second, our research offers practical lessons for survey researchers and pollsters implementing list experiments. We introduced two cost-efficient modifications to the list-experiment procedure that reduce list experiment complexity and difficulty. One of our modifications, the cartoons, is designed with less literate populations in mind. Our other modification, private tabulation, could be also be used in settings where literacy rates are high. The general lesson is that the list experiment should be implemented in a way that minimizes the cognitive burden on respondents. A related implication is that users of the list experiment must pay careful attention not just to the various design considerations emphasized in the methodological literature, but also to the concrete details of implementation.

[15] Simpser (2017)'s statement regards the list experiment ("item-count technique"), as well as randomized response technique and the "Asking about others" approach.

Third, this study has demonstrated the utility of including completely non-sensitive list experiments and corresponding direct questions in a study. These non-sensitive list experiments can be used for participants to practice and gain experience with the question format, before moving on to the real list experiment of substantive interest. They can also be used to identify sub-groups in the study sample that are likely to have particular difficulty with the list experiment format. We advocate for researchers and practitioners employing the list experiment to include non-sensitive list experiments on surveys using list experiments to study substantive, sensitive topics.

Finally, we wish to reflect on the underlying topic motivating this piece and this issue: survey error. The challenge of sensitive topics and behaviors has long beguiled research that relies on survey data because of strong incentives respondents have to be self-censor. Given the potential costs of divulging illegal actions or unpopular undesirable attitudes, the risks simply are not worth being truthful. When such questions yield lies 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 the list experiments work and whether the complexity and difficulty of the task may introduce new forms of error. Unlike predictable downward biases present with direct questions about sensitive attitudes and behaviors, list experiments may suffer from drivers of response error that are not predictable. Thus, in designing list experiments, survey researchers must consider potential tradeoffs between reducing the costs associated with honesty and increasing the costs associated optimally answering the question.

6 Tables

Table 1: Non-Sensitive 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

Table 2: Conventional List Experiment Estimates

Variable	Value	Direct	LE 1	LE 2
Overall		0.14	0.17**	0.01
Education	Less than primary	0.25	−0.36	0.83*
	Primary	0.17	0.27*	0.02
	Some Secondary	0.22	0.23	−0.14
	Secondary	0.10	0.26**	−0.06
	Tertiary	0.08	−0.14	0.20*
Numeracy	0-4 Activities	0.16	0.13	−0.30**
(Mean=5.2)	5-9 Activities	0.13	0.19*	0.17*

*$p < .10$, **$p < .05$, ***$p < .01$ for Welch's T-test with unequal variance comparing treatment and control lists for each list experiment.

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(12):1630–1660.

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(1):101–14.

Ahlquist, John S., Kenenth R. Mayer and Simon Jackman. 2014. "Alien Abduction and Voter Impersonation in the 2012 US General Election: evidence from a survey list experiment." Working paper.

Biemer, Paul and Gordon Brown. 2005. "Model-Based Estimation of Drug Use Prevalence Using Item Count Data." *Journal of Official Statistics* 21(2):287–308.

Blair, Graeme, C. christine Fair, Neil Malhotra and Jacob N Shapiro. 2014. "Poverty and Support for Militant Politics: Evidence from Pakistan." *American Journal of Political Science* 57(1):30–48.

Blair, Graeme and Kosuke Imai. 2012. "Staistical Analysis of List Experiments." *Political Analysis* 20(1):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(511):1304–1319.

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

Corstange, D. 2010. “Vote Buying under Competition and Monopsony: Evidence from a List Experiment in Lebanon.” Presented at the Annual Meeting of the American Political Science Association.

de Jonge, C.K. and D.W. Nickerson. 2014. “Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys.” *Political Behavior* 36(3):659–682.

de Smedt, Johan. 2009. “‘No Raila, No Peace!’ Big Man Politics and Election Violence at the Kibera Grassroots.” *African Affairs* 108(433):581–598.

Droitcour, J., R.A. Caspar, M.L. Hubbard, T.L. Parsley, W. Visscher and T.M. Ezzati. 1991. “The item count technique as a method of indirect questioning: A review of its development and a case study application.” *Measurement Errors in Surveys* pp. 185–210.

Flavin, P. and M. Keane. 2009. “How angry am I? Let me count the ways: Question format bias in list experiments.” *Unpublished Manuscript* pp. 1–17.

Glynn, A.N. 2013. “What Can We Learn With Statistical Truth Serum?: Design and Analysis of the List Experiment.” *Public Opinion Quarterly* 77(1):159–172.

Gonzalez-Ocantos, E., C.K. de Jonge, C. Meléndez, J. Osorio and D.W. Nickerson. 2011. “Vote buying and social desirability bias: Experimental evidence from Nicaragua.” *American Journal of Political Science* 56(1):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(1):37–67.

Humphreys, Macartan and Jeremy Weinstein. 2006. “Handling and Manhandling Citizens in Civil War.” *American Politcal Science Review* 100:429–447.

Kagwanja, Peter Mwangi. 2003. "Facing Mount Kenya or Facing Mecca? The Mungiki, Ethnic Violence, and the Politics of the Moi Succession in Kenya 1987-2002." *African Affairs* 102(406):25–49.

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.

Kanyinga, Karuti and James D. Long. 2012. "The Political Economy of Reforms in Kenya: The Post-2007 Election Violence and a New Constitutionalism." *African Studies Review* 55(1):31–51.

Kasara, Kimuli. 2013. "Separate and Suspicious: Local social and political context and ethnic tolerance kin kenya." *Journal of Politics* 75(4):921–936.

Krosnick, Jon A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." *Journal of Cogitive Psychology* 5:213–36.

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

LaBrie, J.W. and M 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(2):510–533.

Lyall, Jason, Graeme Blair and Kosuke Imai. 2013. "2013Explaining Support for Combattants During Wartime: A Survey Experiment in Afghanistan." *American Political Science Review* 107(4):679–705.

Prior, Markus. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *Journal of Politics* 71(3):893–908.

Rosenfeld, Bryn, Kosuke Imai and Jacob N Shapiro. 2016. "An Empirical Validation Study of Popular Survey Metholdogies for Sensitive Questions." *American Journal of Political Science* 60(3):783–802.

Scacco, Alexandra. 2016. "Anatomy of Riot: Participation in Ethnic Violence in Nigeria." Unpublished Book Manuscript.

Simpser, Alberto. 2017. "Lying about Cheating: A Theory of Underreporting Sensitive Information." Working paper.

Singer, Elanor, Hans-Jurgen Hippler and Norbert Schwarz. 1992. "Confidentiality assurances in surveys: Reassurance or threat?" *International Journal of Public Opinion Research* 4(3):256–268.

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.

Tourangeau, Robert. 1984. Cognitive Sciences and Survey Methods. In *Cognitive aspects of survey methodology: building a bridge between disciplines*, ed. T. Jabine, M. Straf, J. Tanur and Robert Tourangeau. Washington, DC: National Academy Press pp. 73–199.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5):859–93.

Tsuchiya, Takahiro, Yoko Hirai and S. Ono. 2007. "A study of the properties of the item count technique." *Public opinion quarterly* 71(2):253–272.

Zigerelli, L.J. 2011. “You Wouldn’t Like Me When I’m Angry: List Experiment Misreporting.” *Social Science Quarterly* 92(2):552–62.