# Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan*

Graeme Blair† Kosuke Imai‡ Jason Lyall§

Forthcoming in *American Journal of Political Science*

**Abstract**

List and endorsement experiments are becoming increasingly popular among social scientists as indirect survey techniques for sensitive questions. When studying issues such as racial prejudice and support for militant groups, these survey methodologies may improve the validity of measurements by reducing non-response and social desirability biases. We develop a statistical test and multivariate regression models for comparing and combining the results from list and endorsement experiments. We demonstrate that when carefully designed and analyzed, the two survey experiments can produce substantively similar empirical findings. Such agreement is shown to be possible even when these experiments are applied to one of the most challenging research environments: contemporary Afghanistan. We find that both experiments uncover similar patterns of support for the International Security Assistance Force among Pashtun respondents. Our findings suggest that multiple measurement strategies can enhance the credibility of empirical conclusions. Open-source software is available for implementing the proposed methods.

**Key Words:** indirect questioning; item count technique; list experiments; sensitive questions; endorsement experiments; survey experiments; conflict settings

---

*The proposed methods can be implemented via an R package, list: Statistical Methods for the Item Count Technique and List Experiment (Blair and Imai, 2011), which is freely available at the Comprehensive R Archive Network (`http://cran.r-project.org/package=list`). Replication data is available at `http://hdl.handle.net/1902.1/21243`. Financial support for the survey from Yale's Institute for Social and Policy Studies's Field Experiment Initiative and the Macmillan Center for International and Area Studies is gratefully acknowledged. Additional support from the Air Force Office of Scientific Research (Lyall; Grant FA9550-09-1-0314) and the National Science Foundation (Imai; Grant SES–0849715) is also acknowledged. This research was approved by Yale's Human Subjects Committee under IRB protocol #1006006952. We thank Yuki Shiraito for his methodological advice and Aila Matanock, Justin Phillips, and the seminar participants at Kyusyu University, Princeton University, and the University of Michigan for helpful comments. The editor and three anonymous reviewers provided useful suggestions.

†Ph.D. candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: gblair@princeton.edu, URL: `http://graemeblair.com`

‡Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: `http://imai.princeton.edu`

§Associate Professor of Political Science, Department of Political Science, Yale University, New Haven, CT 06520. Phone: 203–432–5264, Email: jason.lyall@yale.edu, URL: `http://www.jasonlyall.com`

# 1 Introduction

Eliciting truthful responses to sensitive survey questions is one of the major methodological challenges in modern social science research. Asking direct questions about such issues as racial prejudice and support for militant groups can lead to significant biases due to dishonest responses and refusals. Even seemingly innocuous questions about turnout in an election are known to result in unreliable answers due to social desirability bias. In some research environments, notably conflict settings, reliance on direct questions may even endanger enumerators and respondents in addition to yielding inaccurate measurements.

In recent years, list and endorsement experiments have become increasingly popular as survey methodology to overcome this measurement problem (see for example references in Tsuchiya, Hirai, and Ono, 2007; Corstange, 2009; Holbrook and Krosnick, 2010; Glynn, 2013; Bullock, Imai, and Shapiro, 2011; Gonzalez-Ocantos et al., 2012; Blair and Imai, 2012).[1] These survey experiments represent indirect questioning techniques in which the individual responses to sensitive questions are not directly revealed. List experiments, also known as the item count technique, use aggregation: respondents are asked to count the number of items on a list which includes a sensitive item (Raghavarao and Federer, 1979; Miller, 1984; Kuklinski, Cobb, and Gilens, 1997). By contrast, endorsement experiments rely upon subtle cues: respondents are asked to rate their support for policies endorsed by socially sensitive actors (Blair et al., 2013; Bullock, Imai, and Shapiro, 2011; Lyall, Blair, and Imai, 2013). New statistical methods have been developed so that researchers can conduct multivariate regression analysis for each of these survey techniques (Corstange, 2009; Glynn, 2013; Bullock, Imai, and Shapiro, 2011; Imai, 2011; Blair and Imai, 2012).

In this paper, we develop a statistical test and multivariate regression models for comparing and combining the results from list and endorsement experiments. One of the fundamental principles of survey methodology is the importance of validating measurements with multiple instruments. We argue that this principle is even more crucial when measuring responses to sensitive questions. While

[1] Another popular survey methodology is randomized response technique (Warner, 1965), which is not studied here (see also Gingerich, 2010).

they may reduce non-response and social desirability biases, indirect questioning techniques are often susceptible to measurement error and are affected by the details of implementation (e.g., Flavin and Keane, 2010). In particular, the results of list and endorsement experiments may depend upon the choice of control items and policy questions, respectively, and sometimes yield substantial design effects which complicate inference. Comparing the results from list and endorsement experiments can serve as an effective diagnostic tool and significantly enhance the credibility of empirical findings.

We first develop a statistical test and a graphical method for directly comparing the responses to list and endorsement experiments. The proposed methodology offers a simple way to examine whether these two survey experiments are measuring the same underlying concept. Next, we demonstrate how to compare list and endorsement experiments in the context of multivariate regression analysis. To do this, we show that the models used for list and endorsement experiments are implicitly linked by a latent level of support for the actor or issue in question. We then demonstrate how to estimate the proportion of supporters with data from either survey experiment technique. This enables researchers to investigate whether or not two survey experiments uncover similar relationships between respondents' characteristics and their support level for the sensitive actor or issue.

Finally, while indirect questioning techniques may reduce bias in response, by construction they elicit less information than direct questioning and thus tend to result in inefficient estimates. We demonstrate how to partially recoup this loss of efficiency by combining the different measurements from both list and endorsement experiments. The method is based on the idea that the same underlying quantity can be measured by these two survey experiment techniques to produce more precise estimates under a single statistical model.

We apply the proposed methods to list and endorsement experiments conducted in an extremely challenging environment, namely, contemporary Afghanistan (Lyall, Blair, and Imai, 2013). Specifically, we motivate this paper by tackling the substantively important issue of measuring support for the International Security Assistance Force (ISAF), the NATO-led mission in Afghanistan currently embroiled in a decade-long effort to create a stable Afghan government while defeating an

entrenched insurgency.

This task is made difficult for several reasons. First, we are attempting to measure support for an organization that, while spending billions in an attempt to sway "hearts and minds," is inherently viewed as an occupying army in the eyes of many Afghans. This situation greatly complicates efforts to measure civilian attitudes since there are such clear incentives to either dissemble in an attempt to continue to receive ISAF-distributed assistance or, conversely, to suppress pro-ISAF sentiment to avoid risking backlash from neighbors or insurgent organizations. Second, we conducted the survey within Pashtun-dominated provinces and districts, the very heart of support for the Taliban. Third, and perhaps unsurprisingly, these areas are highly violent, creating logistical issues while raising concerns about both enumerator and respondent safety. Finally, for cultural reasons, respondents were interviewed publicly, creating additional incentive to answer questions instrumentally. Taken together, these issues all point to the pressing need to embrace innovative measurement strategies to minimize biases and to seek more accurate estimates across multiple survey instruments.

While our empirical example is support for ISAF in Afghanistan, the proposed diagnostic tests and statistical models have widespread applicability across multiple issue areas and subfields. The measurement of prejudice toward minorities and immigrants, for example, is one obvious issue where combining different experimental approaches could yield important insights. Similarly, tracking perceptions of corruption across different political actors, parties, or government agencies with multiple experiments could improve estimates of a notoriously difficult concept to measure. A similar strategy can be applied to the challenges of measuring partisan and coethnic bias in the study of voting behavior and collective goods provision. Clearly, these examples are not exhaustive. Instead, they highlight the wide applicability of our proposed methodology.

The rest of the paper is organized as follows. In Section 2, we briefly describe list and endorsement experiments as survey methodologies for eliciting truthful answers to sensitive questions. We also explain how these survey experiments were conducted in Afghanistan. In Section 3, we propose new methodologies for comparing and combining the results from list and endorsement experiments. In particular, we first describe a statistical test and its associated graphical method for assessing the

compatibility of the two survey measurements. We then develop a multivariate statistical model for combining them. We present the results of our multivariate analysis in Section 4. Finally, Section 5 concludes with practical suggestions for applied researchers.

# 2 List and Endorsement Experiments

In this section, we briefly explain list and endorsement experiments using our Afghanistan application as the running example. We demonstrate that these two survey methodologies can be designed to measure the same underlying concept, here, support for ISAF. For additional discussion about the design of list and endorsement experiments, we refer readers to Blair and Imai (2012) and Bullock, Imai, and Shapiro (2011), respectively. Details about our survey experiment are provided in Lyall, Blair, and Imai (2013).

## 2.1 The List Experiment

The standard design for list experiments randomizes a sample of respondents into two groups where a list of several control items is presented to one group (the "control" group) and a list of the same control items plus one sensitive item of interest is read to the other group (the "treatment" group). Respondents are then asked to count the number of items on their list that fit certain criteria. In our Afghanistan application, we asked the following question to the control group,

> `I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.`
>
> `Karzai Government; National Solidarity Program; Local Farmers`

For the treatment group, the same question was read except that the list contained an additional sensitive item referring to ISAF,

> `Karzai Government; National Solidarity Program; Local Farmers; Foreign Forces`

The core idea behind list experiments is that respondents do not directly answer the sensitive question. Rather, they only need to provide enumerators with the aggregate count of items on a list, which also contains control items. This protection of privacy is designed to increase respondents' willingness to provide truthful answers to sensitive survey questions. In the Afghan application, the list experiment was enthusiastically embraced by our enumerators because it shielded the question's intent, thus reducing concern about asking a sensitive question in villages that were either hotly contested or principally controlled by the Taliban.[2]

As a result, no respondent in our survey refused to answer the question or chose "Don't Know" even though both were presented as options. This compares favorably with ISAF's own survey effort, the Afghan National Quarterly Assessment Report (ANQAR), which relies on direct questions to assess attitudes on a range of sensitive topics, including corruption and support of the Afghan government, ISAF, and insurgent groups. In a recent ANQAR wave conducted in November-December 2011, for example, nearly 50% of respondents refused to participate when approached by an enumerator.[3] An additional subset of individuals who participated were later dropped from the results because they had refused to answer or responded "Don't Know" too many times (the exact threshold for removal is classified). Given these refusal rates, along with a reliance on direct questions asked in public settings, it is likely that these responses are shaded by social desirability bias and outright preference falsification.

However, list experiments also have a known major limitation: respondents in the treatment group will reveal their response to the sensitive item if they choose either all items or none. That is, if a respondent answers "support all" ("support none") then we know that he/she supports (does not support) ISAF. As a consequence, respondents may avoid choosing these extreme answers and provide dishonest responses. Such ceiling and floor effects may be unavoidable, especially when the additional item of interest is highly sensitive. Furthermore, the difficulty of ceiling and floor

[2]More generally, indirect questions can also reduce selection bias in choice of sampling locations. Indirect questions may fly under the radar of village gate-keepers, for example, opening up more sites for potential sampling. In addition, indirect questions reduce incentives for enumerators to falsify data, a practice that can arise when a selected site is too dangerous to ask direct questions safely.

[3]The combined refusal to participate and non-contact rate for this ANQAR wave (Wave 14) in our five provinces was: Helmand (52%), Khost (58%), Kunar (14%), Logar (24%), and Uruzgan (79%).

| response value | Control Group | | ISAF Treatment Group | |
|---|---|---|---|---|
| | frequency | proportion | frequency | proportion |
| 0 | 188 | 20.5% | 174 | 19.0% |
| 1 | 265 | 28.9 | 278 | 30.3 |
| 2 | 265 | 28.9 | 260 | 28.3 |
| 3 | 200 | 21.8 | 182 | 19.8 |
| 4 | | | 24 | 2.6 |
| Total | 918 | | 918 | |

Table 1: Observed Data from the List Experiment to Measure Support for ISAF. The table displays the number of respondents for each value of the observed response and its proportions, separately for the control and treatment group. The proportions do not sum to 100% due to rounding.

effects is that they cannot be easily detected and require an additional assumption for statistical adjustment.

Table 1 provides the summary of responses from the Afghan list experiment. We first conduct a statistical test for design effects (Blair and Imai, 2012), which gives the $p$-value of 1. While this result suggests that there is no clear evidence for design effects, the lack of statistical significance does not necessarily imply the absence of design effects. In particular, the test has no or little statistical power for certain design effect magnitudes and proportions of sensitive item supporters. Given the sensitivity of the question and the public nature of the interview, it is important to use another measure for validating the results based on this list experiment.

In sum, while list experiments provide one means of indirectly asking sensitive survey questions, we need external validation in order to know whether the resulting measurements are reliable. We propose to conduct such validation by comparing the results of list experiments with those of endorsement experiments. Before describing how to compare the two survey experiments, we briefly explain endorsement experiments and their application in Afghanistan.

## 2.2 The Endorsement Experiment

Endorsement experiments offer another indirect way of asking sensitive questions than list experiments. Like list experiments, a sample of respondents is randomly divided into two groups. In the control group, respondents are asked to rate the level of their support for a particular policy. For those in the treatment group, the same question is asked except that the policy is said to be

endorsed by an actor of interest. The main idea is to take advantage of subtle cues induced by endorsements (or names) and interpret the difference in responses between the treatment and control groups as evidence of support (or lack thereof) for this actor of interest. Typically, multiple policies are selected so that the measurement does not rely on a single instrument and statistical power is increased by analyzing them together.

In our Afghan context, four policies were selected to measure support for ISAF relative to a control that lacked a specific endorser. These policies were all publicly endorsed and advocated by ISAF across multiple media and were viewed as a state-building "bundle" designed to strengthen the Afghan government collectively. The selected policies include: the direct election of district councils, a practice enshrined in the Afghan constitution but to date ignored; reform of overcrowded prisons, where squalid conditions are routinely denounced by citizens and watch-dog non-governmental organizations alike; reform of the Independent Election Committee, which is tasked with ensuring electoral transparency; and strengthening of the Office of Oversight for Anti-Corruption, which leads investigations into corruption among government and military officials.

We provide an example script from the prison reform question below:

- Control Condition: A recent proposal calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

- Treatment Condition: A recent proposal by ISAF calls for the sweeping reform of the Afghan prison system, including the construction of new prisons in every district to help alleviate overcrowding in existing facilities. Though expensive, new programs for inmates would also be offered, and new judges and prosecutors would be trained. How do you feel about this proposal?

Respondents were asked to indicate their level of support for this proposal (and all others) on a five-point scale: strongly agree, somewhat agree, indifferent, somewhat disagree, and strongly disagree. As in the list experiment, they also had the options of "Refuse to Answer" and "Don't Know."

Two points bear emphasis. First, our proposed statistical models (detailed below) pool responses to these four questions for estimating ISAF support. This methodology assumes, however, that these policies are actually comparable and occupy the same policy domain. We believe that this is the case. All four proposals are domestic public policy programs that center on the Afghan government's ability to address the rampant inefficiencies that currently undermine its legitimacy. It is possible to empirically test the validity of this assumption. Indeed, our analysis reveals that a fifth endorsement question — one asking about support for ISAF's withdrawal in 2014 — clearly occupied a different (foreign) policy domain when compared to the other four policies.[4]

Second, we also assume that these policies provide an opportunity to measure an individual's support for a combatant. Indeed, these policies have been explicitly endorsed by all combatants in the current war in Afghanistan. While seemingly technocratic, the questions also tap into latent support for different combatants since they address wider questions about the goals of the warring parties and the nature of the current Afghan state — and the one that will emerge from the war. The importance of these issues, as well as the widespread nature of each combatant's endorsement of them, is underscored by the low rate of refusal and "Don't Know" each question obtained.[5] Finally, qualitative evidence from initial focus groups and enumerator debriefings suggest that respondents treated the ISAF endorsement as equally credible as the control endorsement. All told, this evidence indicates that these questions are comparable and useful vehicles for measuring our latent trait of interest among respondents.

Figure 1 presents the overall and province-by-province distributions of responses from the endorsement experiment in comparison with those from the list experiment (left column). While there were no "Don't know" (DK) and "Refuse to answer" responses for the list experiment, the endorsement experiment encountered some DKs and refusals, especially in two districts of Helmand province, where our enumerators ran into open fighting. This suggests that those policies them-

[4]When fitting the model with all five questions, the discrimination parameter, which represents the amount of information each question contributes to the final estimate (see Section 3.2.2 for details), is 1.38 for the ISAF withdrawal question. The same parameter takes the values of 0.017, 0.014, 0.014, and 0.014 for the direct election, prison reform, Independent Election Commission, and anti-corruption questions, respectively.

[5]The combined refusal and "Don't Know" for each question was 5.5%, 4.5%, 7.9% and 3% for questions about direct elections, prison reform, the Independent Election Commission, and corruption reform, respectively.

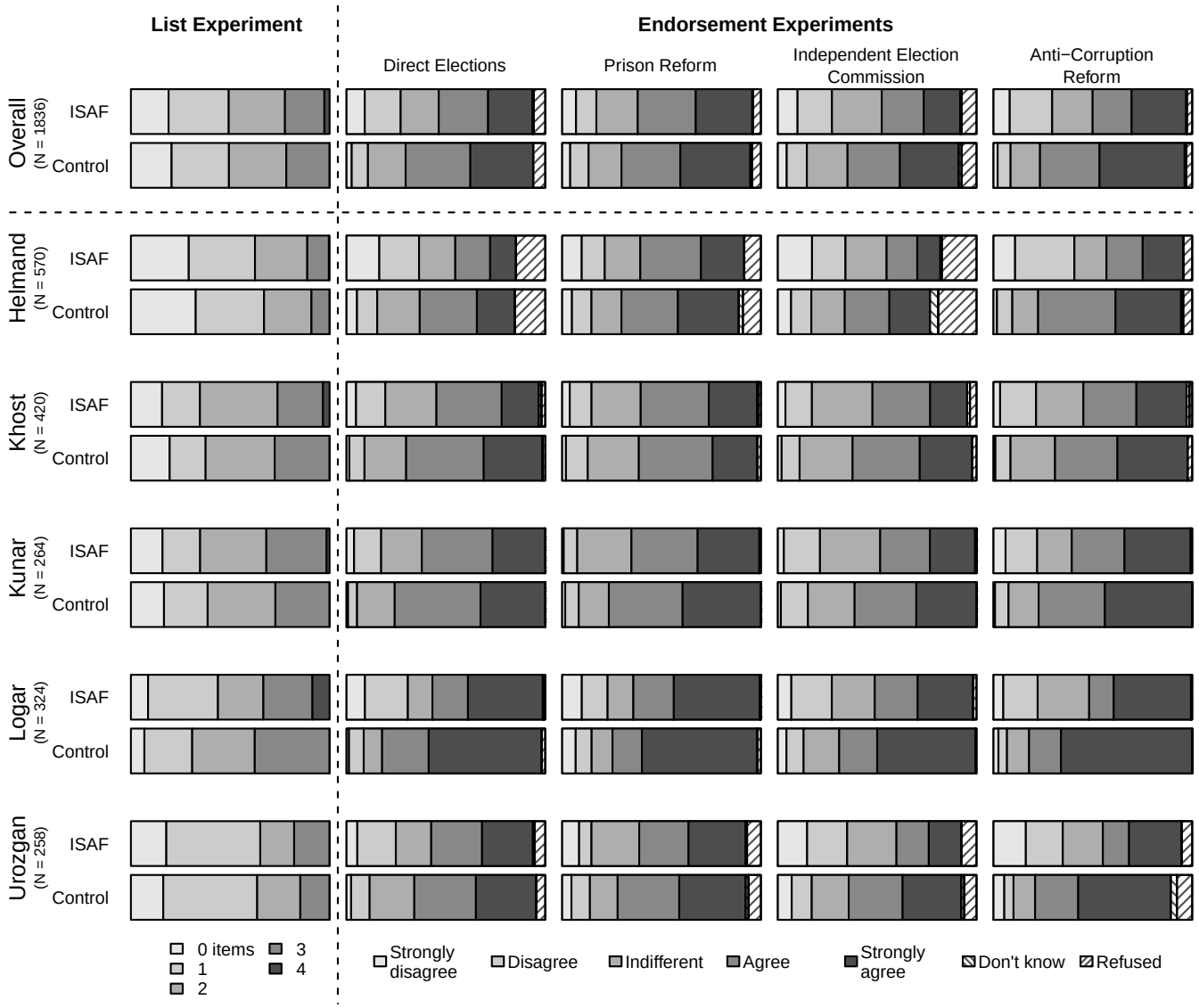


Figure 1: Overall and Provice-by-Province Distributions of Responses from the List and Endorsement Experiments in Afghanistan. In the left column, the plot depicts the distribution of responses to the list experiment questions whereas the remainder of the columns plot the distribution of responses to four policy questions for the endorsement experiment. The overall distribution is given in the top row while the other rows present province-by-province distributions. There were no "Don't know" or "Refuse to answer" for the list experiment questions.

selves may be sensitive in this area. However, the overall non-response and refusal rate is still very low (5.8%) when compared to direct questions used in other surveys. In addition, we observe substantial spatial variation of endorsement effects across provinces whereas the pattern is less clear in the list experiment. In Lyall, Blair, and Imai (2013),we present qualitative evidence that the spatial variation observed in the endorsement experiments is largely consistent with conditions on the ground in each province. By comparison, variation across provinces appears to be more subtle for the list experiment.

While these graphical methods are informative, simply visually comparing the list and endorsement experiment results, as done in Figure 1, does not provide rigorous evidence that these methods are providing comparable measurements of support for ISAF. In the next section, we therefore propose new statistical methods to compare and combine these results.

# 3 The Proposed Methodology

In this section, we begin by introducing a simple statistical test and an associated graphical tool for examining the compatibility of two survey measures. We then show how to compare and combine list and endorsement experiments in a multivariate statistical analysis framework. The key insight is the fact that the same latent support level variable can be used in the statistical models for both types of experiments.

## 3.1 A Statistical Test and Graphical Method

Our analysis begins with a simple graphical method for examining the agreement of measures based on our survey experiments. This method is applicable when each respondent answers both the list and endorsement questions under the same treatment condition, as was the case in our survey. This requires that treatment randomization is conducted across respondents, not within each individual, so that the same respondent is assigned to either the treatment or control group for both experiments.[6] If interference between survey experiments is a concern (e.g., Gaines, Kuklinski, and Quirk, 2007; Transue, Lee, and Aldrich, 2009), at the minimum one should randomize the order of the experiments. We adopted this approach in our pretests but found little evidence for the existence of interference. If researchers decide to conduct list and endorsement experiments on two separate random samples of respondents to avoid interference, the method proposed here will not apply. Note, however, that the statistical method described in Section 3.2 can be employed even under these conditions.

[6] Technically, the same method can be applied to a subset of list and endorsement questions for which the treatment condition is identical even when the treatment is randomized within each respondent. However, this will reduce the statistical power of the proposed analysis.

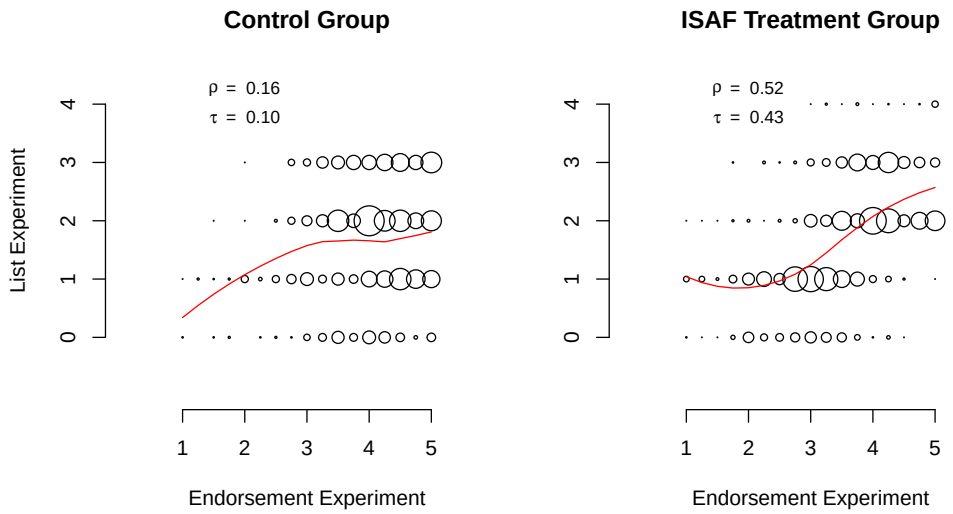


Figure 2: Comparison of Responses from List and Endorsement Experiments in the Afghanistan Survey. For the endorsement experiment (horizontal axis), we plot the mean numerical response on a five point scale across the four endorsement questions. For both experiments, a higher numerical response represents greater support for International Security Assistance Force (ISAF). In the left panel, we present responses in the control group, and in the right panel we present responses in the treatment group for ISAF. The size of open circles is proportional to the number of respondents who gave the corresponding responses. In addition, the lowess curve is presented (red solid line). The Pearson's correlation and Kendall's rank correlation between the two survey measures is represented by "$\rho$" and "$\tau$," respectively. The association between the two measurements is stronger under the treatment condition. We reject the null hypothesis of equality between the two correlation coefficients with the one-sided $p$-value less than 0.001 for both Pearson's and Kendall's correlations.

Figure 2 applies the proposed graphical method to our data. Specifically, we plot each "treated" respondent's answer from the list experiment against his/her average numerical response from the endorsement experiment (right panel) and compare it with the same plot for the control group (left panel). While converting the responses to the endorsement experiment into, say, a five point scale makes an unrealistic assumption that these categories are equally spaced, this simple graphical analysis can give researchers a rough idea about how the treatment can strengthen the association between the two measures. All else equal, if a respondent supports ISAF, their numerical response under the treatment condition for both list and endorsement experiments should be greater than under the control condition. This implies that the association between the two measures should be greater for the treatment group than for the control group.

The right (left) panel of Figure 2 plots the response from the list experiment against the average numerical response from the endorsement experiment among those who are in the ISAF treatment

(control) group. The size of open circle is proportional to the number of respondents, and the lowess curve is also plotted (red solid line). Indeed, the figure reveals that the two measures have a much higher correlation ($\rho = 0.52$) under the treatment condition than under the control condition ($\rho = 0.16$). In addition to Pearson's correlation coefficient, we also compute Kendall's rank correlation (a.k.a. Kendall's $\tau$), which is a nonparametric measure of association. The result is essentially identical in that the correlation is much greater ($\tau = 0.43$) under the treatment condition than under the control condition ($\tau = 0.10$).

We can statistically test the equality of these correlation coefficients. For Pearson's correlation coefficients, it is customary to use the two-sample $z$-test after applying Fisher's $Z$ transformation to each of the two sample correlations (Fisher, 1915). However, it is known that this approximation can be poor if the distribution of underlying data is far from a bivariate normal distribution (Hawkins, 1989). This is exactly the case here since these responses are categorical variables. Thus, we use the nonparametric bootstrap procedure to conduct a statistical test.[7] When this procedure is applied to the data displayed in Figure 2, we find that the difference between the two correlation coefficients is statistically significant with the one-sided $p$-value less than 0.001. The 95% confidence interval of the difference is relatively narrow, $[0.281, 0.356]$. In addition, we apply the same bootstrap procedure using Kendall's $\tau$. This analysis confirms the result based on Pearson's correlation.

In Figure 3, we conduct a similar comparison between list and endorsement experiment questions by examining each endorsement question separately rather than calculating the average of four endorsement questions. With the exception of prison reform, the correlation under the treatment condition remains substantially higher than under the control condition with small $p$-values from the proposed statistical test indicating statistically significant differences.[8] The fact that under the treatment condition the average response from all endorsement questions exhibit a higher correlation than any each individual question demonstrates that the two survey experiments are likely to be

[7] The null hypothesis is the equality of the correlation coefficient between the treatment and control groups and an alternative hypothesis is that the correlation coefficient is greater under the treatment group than under the control group.

[8] The 95% confidence interval of the differences in Pearson's correlation coefficients are $[0.178, 0.259]$ for direct elections; $[-0.057, 0.026]$ for prison reform; $[0.261, 0.344]$ for election commission reform; and $[0.377, 0.457]$ for corruption reform.

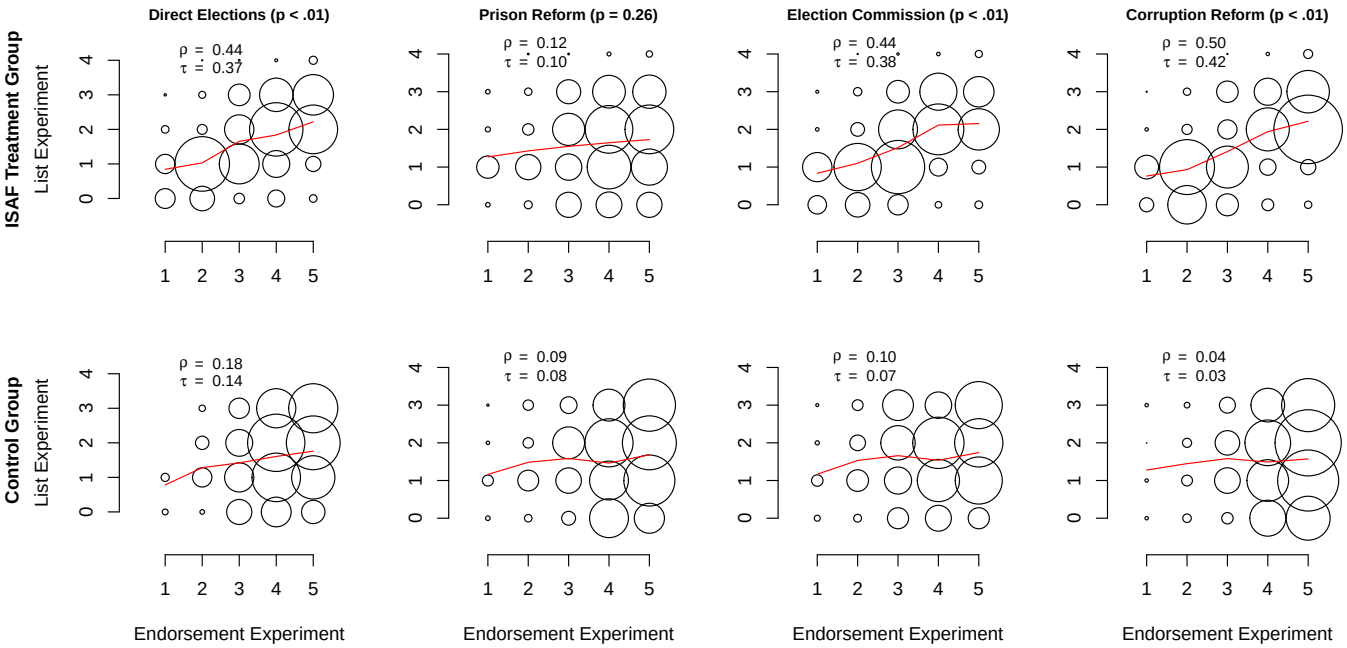


Figure 3: Comparison of Responses from List and Endorsement Experiments in the Afghanistan Survey by Endorsement Question. For the endorsement experiment (horizontal axis), we plot the numerical response on a five point scale. For both experiments, a higher numerical response represents greater support for International Security Assistance Force (ISAF). The Pearson's correlation and Kendall's rank correlation between the two survey measures is represented by "$\rho$" and "$\tau$," respectively. The value given in the parentheses represents the one-sided $p$-value from the statistical test of the equality of two correlation coefficients. In the first row, we present responses in the treatment group for ISAF, and in the second row we present responses in the control group. The size of open circles is proportional to the number of respondents who gave the corresponding responses. In addition, the lowess curve is presented (red solid line).

measuring the same underlying concept – averaging multiple instruments of the same construct typically reduces idiosyncratic noise associated with each survey instrument.

Furthermore, we examine whether this high level of correlation remains when analyzing different subsets of the data defined by key variables theorized in the literature on civil wars to covary with popular support for combatants, notably prior violence (Lyall, Blair, and Imai, 2013; Kalyvas, 2006, pp. 91–104) and the balance of territorial control (Leites and Wolf, 1970; Kalyvas, 2006). Figure 4 presents the results of this analysis. In the left two columns, we present the graphical analyses for subsets of data corresponding to levels of exposure to violence measured by ISAF's Combined Information Data Network Exchange (CIDNE) data on ISAF and insurgent attacks for a one-year period before the survey was conducted. Specifically, the level of violence is measured at

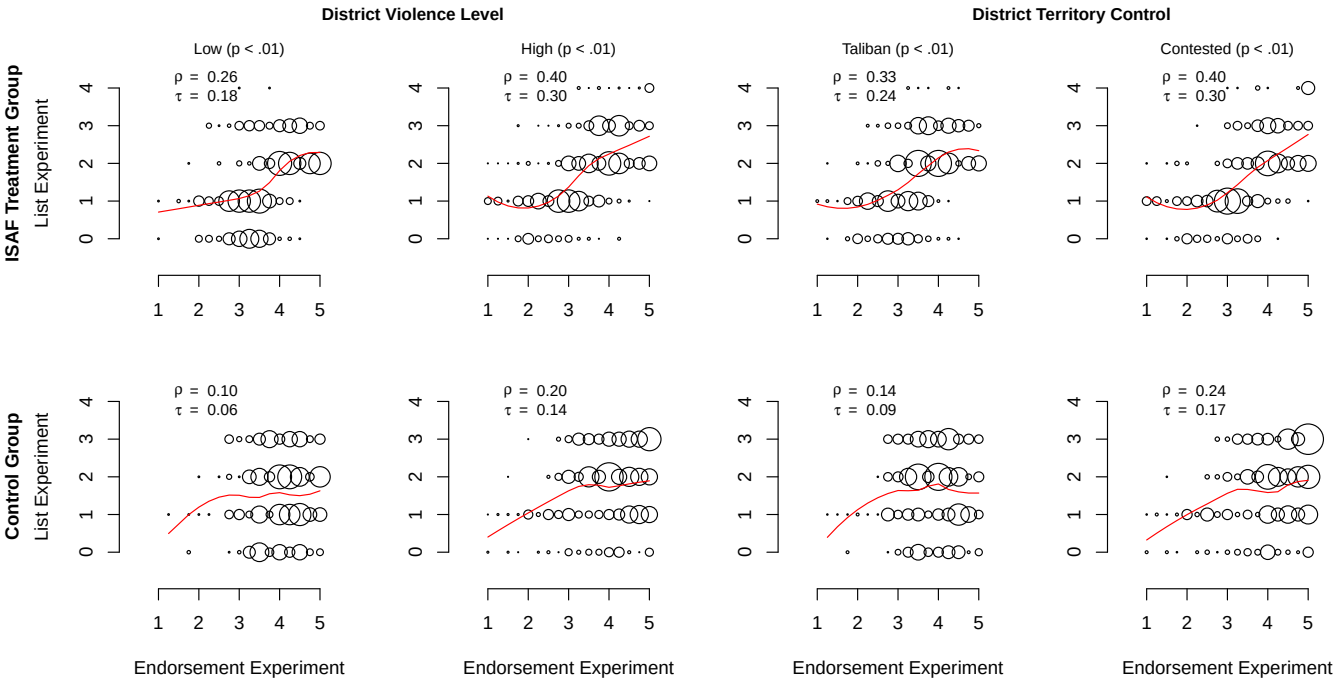


Figure 4: Comparison of Responses from a List Experiment and Four Endorsement Experiments for ISAF by Levels of Violence and Territorial Control. Endorsement experiment measures (horizontal axis) are based on the average numerical responses to all four endorsement questions. In the first row, responses are presented from the ISAF treatment group, and in the second row responses from the control group are presented. The Pearson's correlation and Kendall's rank correlation between the two survey measures is represented by "$\rho$" and "$\tau$," respectively. The one-sided $p$-value based on the statistical test of equality of the two correlation coefficients is presented in the parentheses.

the district level comparing "low violence" (below the mean district violence level in the sample) to "high violence" (above the mean). We also divide the sample into districts that are controlled by the Taliban and those for which control is contested between the Taliban and Afghan/ISAF forces. Across all columns, we find that the correlation between the two measures is much higher under the treatment condition than under the control condition.[9]

In sum, the proposed statistical test and graphical method provide a simple tool to examine whether separate measures from list and endorsement experiments are compatible. Despite the fact that they are based on indirect questions, our ISAF treatment induces a high degree of correlation between the two survey measurements regardless of whether we examine the overall data or only certain theoretically identified subsets. While this provides some confidence that both survey ex-

[9]The confidence intervals of the differences between the correlation coefficients are $[0.314, 0.585]$ for low and $[0.228, 0.420]$ for high violence districts, and $[0.399, 0.634]$ for Taliban and $[0.162, 0.394]$ for contested territorial control.

periments are measuring the same underlying construct of interest, below we propose a statistical methodology that compares and combines these results within a multivariate analysis framework. This framework also incorporates the covariates of respondents for predicting their answers to sensitive questions.

## 3.2 A Statistical Method for Multivariate Analysis

We now more formally compare list and endorsement experiments using multivariate statistical models. The main advantage of this framework is that it allows researchers to directly model the answers to sensitive questions as a function of respondents' characteristics. We also propose a new statistical model that combines the results from list and endorsement experiments. In what follows, we assume that we have a simple random sample of $N$ respondents from a survey in which each respondent answers both list and endorsement experiments. It is also possible to apply the proposed statistical methodology to two separate random samples from the same population if researchers decide to implement list and endorsement experiments to different respondents. Due to space constraints, this extension is not presented in the current paper.

### 3.2.1 A Statistical Model for List Experiments

We begin by briefly reviewing the statistical model for list experiments introduced by Imai (2011) and further developed by Blair and Imai (2012). Let $T_i$ denote the binary treatment assignment variable for respondent $i$, which is equal to 1 if the respondent is assigned to the treatment group and is equal to 0 otherwise. We use $J^L$ to represent the number of control items, where super-script $L$ stands for list experiments. In our survey, respondents with $T_i = 1$ receive a list of four items including ISAF whereas those in the control group are presented with a list of three control items, i.e., $J^L = 3$.

In list experiments, respondents are asked to provide only the total number of items for which their truthful answers are affirmative. For example, a respondent in the control group would provide an integer answer ranging from 0 to $J^L$. We use $Y_i^L$ to represent this aggregate response for each respondent. The distribution of this response variable in our survey is given in Table 1.

Our analysis is based on two assumptions required for standard statistical analyses of list experiments (Imai, 2011; Blair and Imai, 2012). First, the assumption of no design effect states that the addition of a sensitive item does not alter the aggregate response for the control items. Second, the assumption of no liars implies that respondents use truthful answers regarding the sensitive item when giving an aggregate response to the treatment list. Our earlier statistical test suggested no clear evidence for the existence of design effects in the Afghan data (see Section 2.1).[10]

Under this setting, a multivariate statistical model for list experiments can be constructed within the standard likelihood framework. Here, we use a binomial model, though other distributional assumptions are also possible. First, for the control group, we model the observed response as follows,

$$Y_i^L \mid T_i = 0, V_i \;\sim\; \text{Binom}(J^L,\ \text{logit}^{-1}(V_i^\top \psi)), \tag{1}$$

where $V_i$ is a vector of respondents' characteristics including an intercept. For the treatment groups, we use $Y_i^L(0)$ to denote the potential response under the control condition, which under the aforementioned assumptions can be linked to the observed response $Y_i^L$ as follows,

$$Y_i^L \;=\; Y_i^L(0) + Z_i, \tag{2}$$

where $Z_i$ is the latent binary response to the sensitive item.

Then, we model the joint distribution of the latent responses to the control items and the sensitive item among the treated respondents using the following binomial probit regression (though again other distributional assumptions and link functions are also possible),

$$\Pr(Z_i = 1 \mid T_i = 1, V_i) \;=\; \Phi(V_i^\top \gamma), \tag{3}$$

$$Y_i^L(0) \mid T_i = 1, Z_i = z, V_i \;\overset{\text{indep.}}{\sim}\; \text{Binom}(J^L,\ \text{logit}^{-1}(V_i^\top \psi)) \tag{4}$$

[10] As shown by Blair and Imai (2012), modeling deviations from these assumptions is possible but for the sake of simplicity we maintain them for the reminder of the paper.

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Imai (2011) and Blair and Imai (2012) consider various extensions of this basic model by, for example, allowing the aggregate response to the control items, $Y_i^L(0)$, to explicitly depend on the latent response to the sensitive item, $Z_i$, conditional on $V_i$, whereas for the sake of simplicity we assume conditional independence between $Y_i^L(0)$ and $Z_i$. We note that researchers may also wish to model the over-dispersion as done in Imai (2011) and implemented in the accompanying open-source software.[11] Doing so may decrease the precision of the resulting estimates but yield confidence intervals that better reflect the estimation uncertainty.

For fitting this model, we develop a Bayesian model of list experiments so that this model can subsequently be compared and combined with a statistical model of endorsement experiments, which is also Bayesian. We complete our model by choosing the following standard semi-conjugate prior distribution for the parameters in the above list experiment model,

$$\psi \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\ A_\psi) \tag{5}$$

$$\gamma \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\ A_\gamma) \tag{6}$$

where in our empirical analysis of the Afghan data $A_\psi$ and $A_\gamma$ are equal to a diagonal matrix with variance equal to 9 for each parameter. Appendix A.1 gives the Markov chain Monte Carlo (MCMC) algorithm used to sample from the posterior distribution based on this model.

### 3.2.2 A Statistical Model for Endorsement Experiments

Next, we describe the statistical model for endorsement experiments proposed by Bullock, Imai, and Shapiro (2011). Suppose that we have $J^E$ policy questions in which respondents are asked to rate their support for policy $j$ on an $M_j$ point scale. $E$ stands for endorsement experiment. In our survey, for example, there are $J^E = 4$ questions and $M_j = 5$ for all $j$ (i.e., strongly agree, agree, indifferent, disagree, strongly disagree).

The key modeling idea is to combine responses across multiple policy questions using the frame-

[11] In particular, the beta-binomial model may be substituted for the binomial model.

work of item response theory and obtain a single measure of the level of support on the underlying policy dimension. The key assumption underlying this approach is that the selected policies belong to the single dimension, allowing us to combine responses to these policy questions. As detailed above, we believe that our four policies occupy the same domain and measure the same latent trait.

To formally describe our model, let $Y_{ij}^E$ represent respondent $i$'s answer to the question about policy $j$. We consider a variant of the standard ordered probit item response theory model,

$$\widetilde{Y}_{ij}^E \mid T_i \overset{\text{indep.}}{\sim} \mathcal{N}(\beta_j(x_i + T_i s_{ij}^*) - \alpha_j,\ 1) \tag{7}$$

where $Y_{ij}^E = y$ if $\tau_{yj} < \widetilde{Y}_{ij}^E < \tau_{y+1,j}$ with $\tau_{0j} = -\infty$, $\tau_{1j} = 0$, and $\tau_{M_j+1,j} = \infty$ being cutpoints. Here, $\alpha_j$ parameterizes the average unpopularity of policy $j$, $\beta_j$ represents how much policy $j$ differentiates respondents who have different ideologies in this relevant policy dimension (e.g., pro- and anti-reform respondents), and $x_i$ characterizes the ideological position of respondent $i$ (e.g., how pro-reform respondent $i$ is).

We define the latent level of support as $s_{ij} = s_{ij}^* \cdot \text{sgn}(\beta_j)$ so that, for example, $s_{ij} > 0$ implies that respondent $i$ is more likely to voice support for policy $j$ when endorsed by the group, which we interpret as evidence of respondent $i$'s positive support for the group. In our empirical analysis of the Afghan data, we assume $\beta_j > 0$ using a truncated normal distribution for prior. This choice can be justified by the substance of the endorsement questions: answering them affirmatively means a respondent is supportive of domestic policy reforms.

We model support levels and ideal points as a function of respondent characteristics in the following hierarchical manner,

$$x_i \overset{\text{indep.}}{\sim} \mathcal{N}(V_i^\top \delta,\ 1) \tag{8}$$

$$s_{ij} \overset{\text{indep.}}{\sim} \mathcal{N}(V_i^\top \lambda,\ \omega^2). \tag{9}$$

We complete the model by using the following independent conjugate prior distributions for the

model parameters,

$$(\alpha_j \ \beta_j) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \ B) \tag{10}$$

$$\delta \sim \mathcal{N}(0, \ C) \tag{11}$$

$$\lambda \sim \mathcal{N}(0, D) \tag{12}$$

$$\omega^2 \sim \kappa/\chi^2_\nu \tag{13}$$

In Appendix A.2, we present the MCMC algorithm used to sample from the posterior distribution.

As discussed by Bullock, Imai, and Shapiro (2011), the results of endorsement experiments can be sensitive to the choice of policy questions for two reasons. First, typically we can only include a small number of policy questions in the endorsement experiments for the practical and logistical reasons. This means that the estimates will be inherently sensitive. Second, all policy questions may not contribute equally to the estimation of support level because some questions more informative about respondents' ideology than others (i.e., the discrimination parameter $\beta_j$ may vary across policies). Therefore, it is important to examine the values of discrimination parameters and conduct a sensitivity analysis.

### 3.2.3 Comparing List and Endorsement Experiments

Given the two separate models described above, we now detail how their results can be compared. The key insight is to recognize that the probability of supporting the group of interest given respondents' characteristics $V_i$ can be derived from both models. Specifically, for the list experiment model, this quantity is given by,

$$\Pr(Z_i = 1 \mid V_i) = \Phi(V_i^\top \gamma) \tag{14}$$

which follows directly from the model of the latent response to the sensitive item given in equation (3). Similarly, for the endorsement experiment model, we can calculate the probability that

the support parameter $s_{ij}$ takes a positive value for any given policy question $j$ as,

$$\Pr(s_{ij} > 0 \mid V_i) \quad = \quad \Phi(V_i^\top \lambda^*). \tag{15}$$

where $\lambda^* = \lambda/\omega$ is an identifiable parameter.

Given this relationship between the two models, there are several interesting comparisons one can make after fitting each model separately to the data from list and endorsement experiments. First, we can compare the overall proportion of supporters for the group within a target population based upon the two models. This is done by computing the difference in the sample averages of the above quantities using the observed value of $V_i$ for each respondent. This comparison will enable a formal assessment of the overall agreement between the two survey experiments. Second, it is possible to directly compare the coefficients for respondents' characteristics by computing the difference between $\gamma$ and $\lambda^*$ because these two parameters are on the same scale of the normally distributed latent support variable. This analysis will show whether similar patterns of association between support level and respondents' characteristics can be derived from list and endorsement experiments. Indeed, researchers can go beyond the comparison of coefficients and examine how the estimated probability of support changes as a function of covariates. Finally, we can explore the characteristics of respondents who are likely to give more or less compatible responses for list and endorsement experiments. Such an analysis can be conducted by computing the difference in the probability of supporting the group of interest given certain respondent characteristics.

For all of these comparisons, one can easily compute uncertainty estimates and conduct appropriate statistical tests. Since posterior draws for these quantities are available from each of the respective models, this can be done by simply computing their difference for each posterior draw.

### 3.2.4 Combining List and Endorsement Experiments

Finally, we demonstrate how to combine the list and endorsement experiments using a single statistical model. To do this, we take advantage of the fact that both models are based on the latent support level whose distribution is the standard normal but with different conditional means, $V_i^\top \gamma$

for the list experiment model and $V_i^\top \lambda^*$ for the endorsement experiment model. Thus, in order to combine the list and endorsement experiments, we assume the same data generating process for the latent levels of support in both experiments. Specifically, we assume that latent support levels in both experiments are identically distributed by imposing the restriction $\gamma = \lambda^*$, which directly connects the two models. The resulting model makes it possible to analyze list and endorsement experiments together and yields a single estimate of support level for the group of interest. Appendix A.3 describes the MCMC algorithm we use to fit this combined model.

The main advantage of this strategy is that it provides a single, coherent set of estimates from the two survey experiments and increases the statistical efficiency of the resulting estimates. The validity of this combining strategy can be formally assessed by examining standard model comparison statistics such as Bayes factor and Bayesian information criteria (in addition to simply comparing these two parameters after separately fitting the two models). We note that the proposed combined model necessarily places greater weight on the endorsement experiment rather than the list experiment because the former typically comprises several questions. If one wishes to weight them equally, our model can be extended to partially pool endorsement questions through another level of hierarchy and then combine them with list experiment question. Finally, while not explored here either, it is also possible to impose a partial restriction where only some (but not all) coefficients are assumed to be the same between the two models. Such an approach might be useful if researchers wish to formally model the relationships between respondents' characteristics and the degree of compatibility of the two survey experiments. We leave these and other extensions to future work.

# 4 Results of the Multivariate Statistical Analysis

We begin our multivariate statistical analysis by fitting three separate statistical models. To analyze the endorsement experiment, we use the model described in Section 3.2.2 with the same covariates as the ones used by Lyall, Blair, and Imai (2013).[12] Specifically, the individual-level predictors include whether a respondent was harmed by ISAF and the Taliban (both physical harm

[12]For the sake of simplicity, we do not employ multi-level modeling.

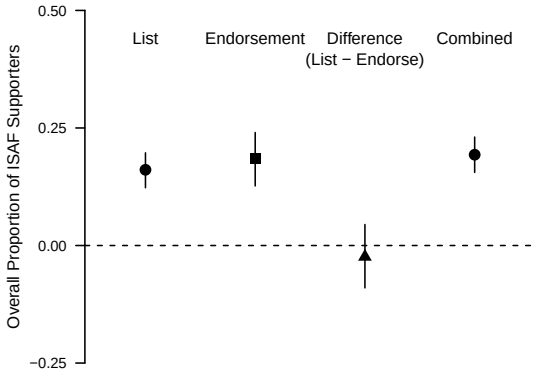


Figure 5: Estimated Overall Proportion of ISAF Supporters based on the List Experiment, Endorsement Experiment, and Combined Models. The difference between the estimates based on the list (first column) and endorsement experiment (second column) models is presented in the third column. The fourth column shows the estimate based on the combined model. The vertical lines represent 95% confidence intervals. The estimates are essentially identical across all three models.

and property damage), whether those who were harmed were subsequently approached by the perpetrator (implying some form of post-harm mitigation efforts), whether a respondent is a member of a Pro-Taliban tribe, how often they encounter ISAF, as well as respondents' age, income, years of education, and years of madrassa schooling. The model also includes several village and district level characteristics. For example, as mentioned in Section 3.1, we include violence count variables and district-level variables recording prior levels of foreign assistance and economic aid.

Similarly, we use the model described in Section 3.2.1 to analyze the list experiment with the same set of covariates. Finally, we conduct a joint analysis of list and endorsement by estimating the combined model described in Section 3.2.4 again with the same set of covariates. The full list of variables are given in Appendix A.4 along with estimated coefficients and standard errors, and Lyall, Blair, and Imai (2013) offer a detailed explanation of each variable. Finally, we use the MCMC algorithms described in the Appendix to fit these models. We run three chains with over-dispersed starting values to monitor convergence (Gelman and Rubin, 1992). After a sufficient degree of convergence is achieved, we take the last half of posterior draws to compute our quantities of interest.

## 4.1 Estimated Overall Levels of Support

We begin with a simple comparison of the estimated overall proportion of those who support ISAF. As Figure 5 demonstrates, the two modes of survey experiment return nearly identical point estimates, with support for ISAF only modest among our respondents (at 16% for the list experiment, 17.5% for the endorsement experiment, and 19% for the combined model), though the 95% confidence interval is somewhat wider for the endorsement experiment. These estimates are obtained by computing the predicted probability of supporting ISAF for each respondent and then averaging this probability across all respondents in the sample. The fact that the estimated overall difference between the two measurement strategies is almost zero is remarkable given the significant differences in question format between list and endorsement experiments. Given the similarity of the results between the list experiment and endorsement experiment models, it is no surprise that the combined model also yields an essentially identical estimate.[13] Taken together, these results support our assumption that the two survey experiments are measuring the same underlying concept.

What is the relative efficiency of these estimates? The width of the 95% confidence interval is 0.124 for the endorsement experiment model, 0.080 for the list experiment model, and 0.075 for the combined model. Thus, in our case, as expected, the combined model provides the most efficient estimate though the efficiency gain relative to the list experiment model is modest. To place these numbers in a perspective, we note that if the direct questioning was used with the same sample size and returned a similar estimate (i.e., 0.2), then the width of the 95% confidence interval would equal 0.037, which is approximately half of that from the combined model.

As discussed at the end of Section 3.2.2, the results of endorsement experiments can be sensitive to the choice of policy questions. Indeed, our estimates of the discrimination parameter differ somewhat across policies with $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) = (0.93, 0.56, 0.74, 0.74)$, giving the largest weight to the direct elections policy question. We further conduct a sensitivity analysis by dropping each policy question and examining its impact on our estimates. The results are presented in

[13] It may appear to be counter-intuitive that the combined estimate is not a weighted average of the list and endorsement estimates, but this could arise due to the non-linearity of the model.

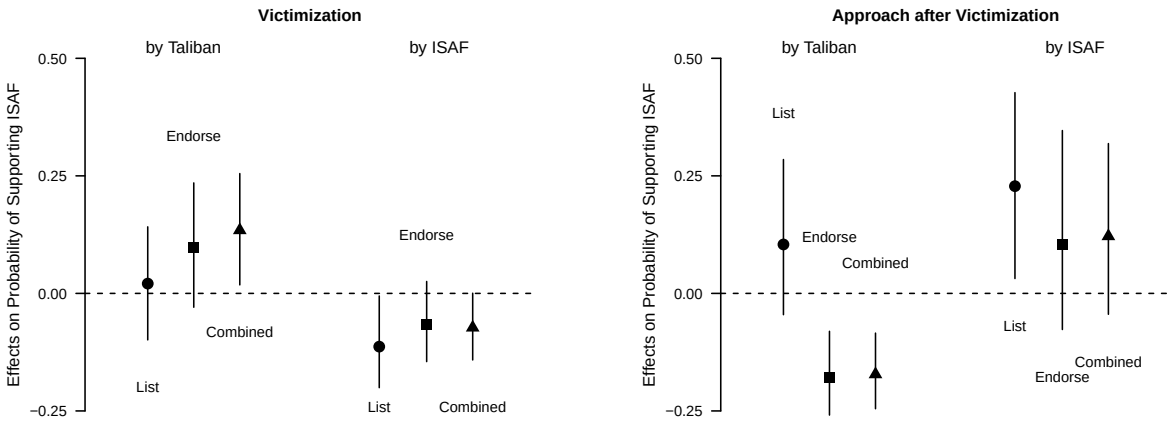


Figure 6: Estimated Average Marginal Effects of Taliban and ISAF Victimization and their Post-harm Mitigation Efforts ("Approach" by Combatants) on the Probability of Supporting ISAF. Three estimates based on the list experiment, endorsement experiment, and combined models are reported along with the difference between the list and endorsement experiment models. The vertical lines represent 95% confidence intervals. Across three models, the victimization and subsequent approach by ISAF have negative and positive effects, respectively.

Appendix A.5. As expected, dropping the direct elections question has the largest impact on our estimate: driving the estimated level of support for ISAF further down. By contrast, dropping any of the other three policies has relatively little impact on our estimate. The sensitivity analysis suggests that the level of support for ISAF might be well lower than what is presented here.

## 4.2 Estimated Marginal Effects on Levels of Support

We next explore the conditional nature of ISAF support by first assessing the marginal effect of self-reported Taliban and ISAF victimization on respondents' support for ISAF (the left panel of Figure 6) across the different survey methods. We find that Taliban victimization leads to a modest increase in support for ISAF. While the combined model estimates a more precise positive effect, the estimates from the list and endorsement experiment models are inconclusive. By contrast, ISAF victimization is associated with consistently negative effect across three models. While Taliban violence may not drive Pashtuns into the arms of ISAF, violence perpetrated by ISAF actually push respondents toward the Taliban. This asymmetrical result is consistent with the theory and empirical findings of Lyall, Blair, and Imai (2013).

We also examine the relationship between support for ISAF and post-harm efforts by the Taliban

and ISAF to mitigate the effects of their violence. Interestingly, we observe that the Taliban's post-harm efforts have an inconsistent effect on attitudes toward ISAF across these models. While the list experiment model estimates a modest positive effect, suggesting that Taliban post-harm efforts may actually be associated with an increase in ISAF support, the endorsement experiment model suggests the opposite, which is also favored by the combined model. By contrast, ISAF post-harm mitigation efforts are associated with an estimated positive effect on ISAF support regardless of which survey experiment we analyze. Furthermore, the combined model provides a similar point estimate but with narrower confidence intervals. As with all of these comparisons, results that are supported by both the list and endorsement experiment models should be viewed as more credible than those that are inconsistent across models.

We next proceed to investigate how key combatant variables are associated with the degree of agreement (or disagreement) on the estimated levels of ISAF support across list and endorsement experiments. In Figure 7, we present the results with respect to two variables identified by the existing literature as determinants of civilian attitudes and actions in civil war: (1) the amount of aid or development funds allocated to a given village or area (Beath, Christia, and Enikolopov, 2011; Berman, Shapiro, and Felter, 2011; U.S. Army, 2007), here measured by district-level Commander's Emergency Response Program, or CERP, spending (top panel) and (2) the level of control (Kalyvas, 2006) exerted by the Taliban and ISAF at the district level in the months before our survey was conducted (bottom panel).

Beginning with CERP spending, we find an intriguing empirical pattern. Specifically, while the endorsement experiment (and combined results, not shown) reveals little effect of CERP spending on the probability of supporting ISAF, the estimates based on the list experiment are strongly positive with a large effect size. While one possible explanation for this list experiment finding is that aid is shifting attitudes in a pro-ISAF direction, we believe that this is unlikely because we do not observe the same increase in the more subtle endorsement experiments. Given that list experiments are more direct in their elicitation approach when compared with endorsement experiments, we conjecture that individuals felt pressured to respond in a positive direction when

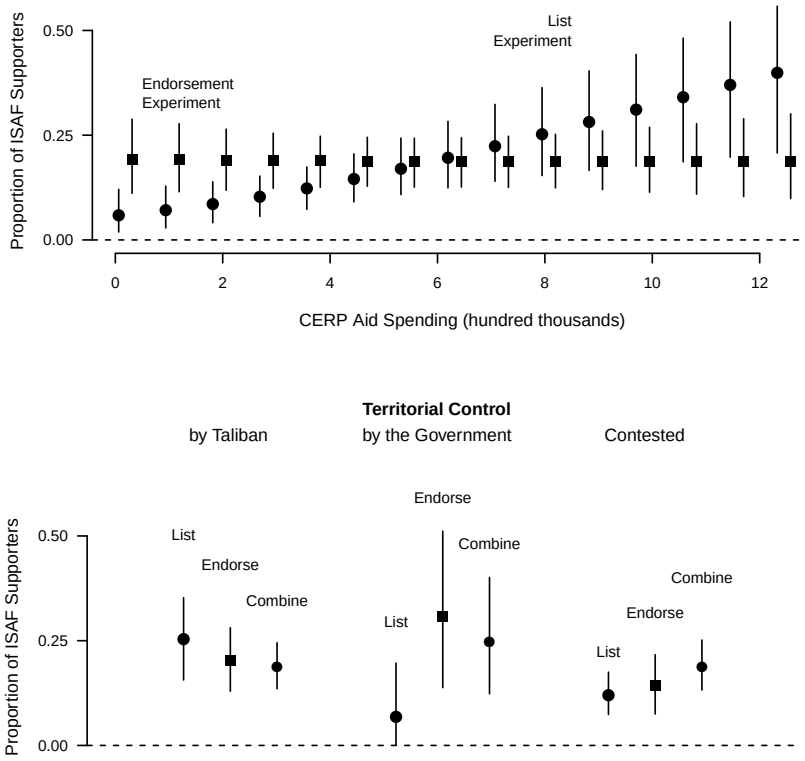


Figure 7: Estimated Proportion of ISAF Supporters based on the List Experiment, Endorsement Experiment, and Combined Models As a Function of by the Amount of Aid (top panel) and the Territorial Control (bottom panel). In the top panel, the proportion of supporters of ISAF were calculated across varying levels of aid spending under the Commander's Emergency Response Program (CERP) based on the endorsement experiment and list experiment models. The estimated proportion of supporters of ISAF within districts designated by ISAF as controlled by the Taliban, government forces, and contested districts were calculated based on the list experiment, endorsement experiment and combined models in the bottom panel. The vertical lines represent 95% confidence intervals.

asked about ISAF, perhaps out of a belief that continued assistance was conditional on obtaining positive responses. This result underscores the need to cross-reference empirical results across multiple survey experiments in order to increase confidence in our findings.

Finally, we may also be concerned that support is endogenous to the relative level of control exercised by the combatants. In our case, it is plausible that the probability of supporting ISAF is highest in ISAF controlled areas, lowest in Taliban-dominated areas, and intermediate in contested areas as fence-sitting individuals hide their views.

We present evidence concerning these hypotheses in the bottom panel of Figure 7 for our 21

districts. Several trends are notable. First, the list and endorsement experiments provide broadly similar estimates of the probability of supporting ISAF within Taliban-dominated and contested districts. These findings underscore the ability of these instruments to solicit opinions on extremely sensitive issues even under difficult conditions. Second, we do observe some divergence across the list and endorsement experiment results for government-controlled districts. We must be cautious in drawing firm conclusions, however, for there are only three government-controlled districts in our sample. Somewhat surprisingly, we find despite concerns over endogeneity and social desirability bias that the lowest estimate for ISAF support is returned by the list experiment in government-controlled districts.

Third, an examination across districts reveals that the estimates from our combined model are similar across Taliban-dominated and contested areas. This is a striking finding, one at odds with expectations in the literature that civilian attitudes toward combatants are purely endogenous to control by combatants. The fact that the combined model yields estimates of the effects on the probability of supporting ISAF that are only a few points higher than the other district types reinforces the fact that the relationship between control and attitudes may be far more complicated than captured by current theoretical work.[14]

In sum, the results presented in this section demonstrate that when carefully designed and analyzed, list and endorsement experiments can generate substantively similar conclusions even when these survey experiments are conducted in a challenging research environment. We found that the list and endorsement experiment models yield essentially identical estimates of the overall proportion of ISAF supporters among Pashtun men. When the results from the two models diverge, as in the case of CERP spending, researchers should interpret findings with caution. On the other hand, confidence in the robustness of our results is increased when the two models converge on similar empirical predictions. The combined model itself yields even more precise estimates of the quantity of interest, suggesting that such an approach could be invaluable in (post-)conflict and

[14]We also divided our population into pro-Taliban and neutral Pashtun tribes to determine if the experiments performed differently for these subsets. The list and endorsement experiments provided similar results within these populations.

other similar settings where measuring attitudes toward sensitive issues or actors often induces noise as well as potential bias.

# 5 Concluding Remarks and Practical Suggestions

List and endorsement experiments are becoming popular tools for eliciting truthful responses to sensitive questions. They have been shown to be effective measurement strategies even in extremely challenging research environments such as in our empirical application: wartime Afghanistan. However, since these survey techniques are based on indirect questioning, it is important for researchers to cross-validate their measurements using different survey instruments. To enable such comparisons, we introduce statistical methods to compare and combine the results from these survey experiments. For the first time, we demonstrate how to directly link data from the two survey experiments through the latent level of support, and then show how to estimate a key quantity of interest — the proportion of supports of the actor or issue under study — using data from either questioning technique. We also develop multivariate techniques in a single statistical framework to estimate marginal effects of covariates on the level of support.

In our empirical application, we find that patterns of estimated support for ISAF among Pashtun men are remarkably consistent across list and endorsement experiments. This finding highlights the promise of the proposed empirical strategy given that the survey was conducted in a difficult research setting marked by insurgent control, high degrees of violence, and persistent efforts by ISAF to sway public attitudes through widespread aid programs. Furthermore, these results underscore the importance of drawing on multiple methods in a systematic fashion to increase confidence that we are accurately measuring the quantity of interest.

We conclude this paper by offering a set of methodological and practical suggestions for researchers seeking to combine these and other indirect survey techniques. General methodological recommendations about list and endorsement experiments are given elsewhere (Bullock, Imai, and Shapiro, 2011; Blair and Imai, 2012). Here, we offer additional suggestions for those who are interested in implementing these survey techniques.

First, we recommend that researchers randomize the treatment across, not within, individuals. That is, a respondent should be randomly assigned to a control or treatment group and then should only receive either all control or all treatment group questions. In cases of multiple treatments, the respondent should only receive questions about one sensitive item. This fusing of a respondent to all control or treatment questions should extend across the survey instruments because it permits direct comparison of the same individual across different survey instruments via our statistical test and its associated graphical method.

While the inability to exploit within-respondent treatment variation may reduce statistical efficiency, this strategy has several additional benefits. Specifically, it avoids triggering suspicion that might arise when repeatedly asking about a range of sensitive items or actors, thus exposing the fact that the survey is seeking comparative answers. It also greatly simplifies logistics by eliminating the need to manage a large number of different survey versions to account for all possible combinations of treatment assignments.

Second, we note that list experiments appear to be more prone to social desirability bias than endorsement experiments. This makes sense because list experiments are more direct than endorsement experiments. In our survey, for example, the list experiment asks, though indirectly, a question about support for ISAF whereas the endorsement experiment question is about support for a particular policy endorsed by ISAF rather than ISAF itself. Thus, list experiments may not be suitable for extremely sensitive questions. We observed this first-hand when we attempted to measure support for the Taliban using a list experiment. Despite having the same control items as our ISAF experiment, the inclusion of such a sensitive item triggered dramatic floor and ceiling effects: no respondent answered either "0" (i.e. support none) or "4" (support all). Individuals clearly strategically hid their support (or lack thereof) for the Taliban within the remaining empirical distribution, making it impossible for us to recover measures of support. As demonstrated in the previous section, the list experiment was also notably sensitive to the amount of aid rendered to a district.

Third, to avoid relying on estimates that are dependent on a single measurement strategy,

we recommend that researchers use multiple questioning techniques. Crucially, researchers should present the estimates based on each different experimental technique to confirm the robustness of the conclusion to varying measurement strategies, or to identify the analyses that are more sensitive to differences in question format.

Fourth, and on a more practical note, we found it essential to engage in multiple pretests and focus groups in the areas to be sampled (rather than, say, a convenience sample in Kabul or its outskirts) to test for sensitivity to question order. The list experiment proved particularly time-intensive, both in terms of training the enumerators and in ensuring that average Afghan respondents who have very little formal education could understand the approach. We preceded our list experiment with a practice example to ensure that respondents understood the mechanics of the design. We also piloted multiple versions of the list and endorsement experiments to identify suitable control items and policy questions in order to avoid floor and ceiling effects and other complications.

This research also suggests several natural extensions. The proposed multiple measurement strategy could be employed in a variety of different (and difficult) empirical settings where violence, social desirability concerns, and the need to assess attitudes on sensitive items collide. On the methodological front, a similar modeling strategy can be used to model other indirect questioning techniques such as randomized response as well as variants of list and endorsement experiments by linking different models via the latent level of support as done in this paper. We can also extend the combined statistical model introduced in this paper to model individual and spatial characteristics that are responsible for driving different responses across these survey experiments. Armed with this knowledge, researchers would be able to customize an array of experimental approaches given the particular challenges they are likely to encounter in their study population.

# A Appendix

In this appendix, we describe the details of the Markov chain Monte Carlo algorithms for three models: the list experiment model, the endorsement experiment model, and the combined model. We also present the estimated coefficient from each of these three models.

## A.1 The List Experiment Model

**Step 1:** Update the coefficients for the control item model, $\psi$, using the random walk Metropolis step where the proposal draw is obtained from the multivariate normal distribution with mean equal to its previous draw and a pre-specified tuning variance parameter $S$. The acceptance ratio is given by,

$$\min\left(1,\ \frac{\prod_{i=1}^{n}\left\{\text{logit}^{-1}\left(V_i^\top\psi^{(t)}\right)\right\}^{Y_i^L(0)^{(t-1)}}\left\{1-\text{logit}^{-1}\left(V_i^\top\psi^{(t)}\right)\right\}^{J^L-Y_i^L(0)^{(t-1)}}\mathcal{N}(\psi^{(t)},A_\psi)}{\prod_{i=1}^{n}\left\{\text{logit}^{-1}\left(V_i^\top\psi^{(t-1)}\right)\right\}^{Y_i^L(0)^{(t-1)}}\left\{1-\text{logit}^{-1}\left(V_i^\top\psi^{(t-1)}\right)\right\}^{J^L-Y_i^L(0)^{(t-1)}}\mathcal{N}(\psi^{(t-1)},A_\psi)}\cdot\frac{\mathcal{N}(\psi^{(t)},S)}{\mathcal{N}(\psi^{(t-1)},S)}\right)$$

where $\mathcal{N}(\cdot,\cdot)$ represents the (possibly multivariate) normal density function.

**Step 2:** Update the coefficients for the sensitive item model, $\gamma$, using the data augmentation algorithm.

$$Z_i^{*(t)} \mid Z_i^{(t-1)},\gamma^{(t-1)} \sim \begin{cases} \mathbf{1}\{Z_i^{*(t)}\geq 0\}\mathcal{N}\left(V_i^\top\gamma^{(t-1)},\ 1\right) & \text{if } Z_i^{(t-1)}=1 \\ \mathbf{1}\{Z_i^{*(t)}<0\}\mathcal{N}\left(V_i^\top\gamma^{(t-1)},\ 1\right) & \text{if } Z_i^{(t-1)}=0 \end{cases} \quad \text{for each } i$$

$$\gamma^{(t)} \mid \left\{Z_i^{*(t)}\right\}_{i=1}^{n} \sim \mathcal{N}\left(\left(\sum_{i=1}^{n}V_iV_i^\top+A_\gamma\right)^{-1}\sum_{i=1}^{n}V_iZ_i^{*(t)},\ \left(\sum_{i=1}^{n}V_iV_i^\top+A_\gamma\right)^{-1}\right)$$

**Step 3:** Sample the latent responses to the sensitive and control items, $(Z_i, Y_i^L(0))$ for units with $T_i=1$ and $1\leq Y_i^L\leq J^L-1$. For units with $Y_i^L=J^L$ or $Y_i^L=0$, we set $Z_i^{(t)}=1$ and $Z_i^{(t)}=0$, respectively. For each unit, we sample the latent response to the sensitive item according to the following distribution,

$$Z_i^{(t)} \mid \psi^{(t)},\gamma^{(t)} \sim \text{Bernoulli}\left(\frac{\text{Binom}\left(Y_i^L-1;J^L,\ \pi_i^{(t)}\right)\cdot\Phi\left(\mu_i^{(t)}\right)}{\text{Binom}\left(Y_i^L-1;J^L,\ \pi_i^{(t)}\right)\cdot\Phi\left(\mu_i^{(t)}\right)+\text{Binom}\left(Y_i^L;J^L,\ \pi_i^{(t)}\right)\cdot\left\{1-\Phi\left(\mu_i^{(t)}\right)\right\}}\right)$$

where $\pi_i^{(t)}=\text{logit}^{-1}(V_i^\top\psi^{(t)})$, $\mu_i^{(t)}=V_i^\top\gamma^{(t)}$, and $\text{Binom}(\cdot,\cdot)$ represents the binomial density function. For the responses to the control items, we set $Y_i^L(0)^{(t)}=Y_i^L-Z_i^{(t)}$.

## A.2 The Endorsement Experiment Model

**Step 1:** Sample the cutpoint parameters, $\{\tau_{lj}^{(t)}\}_{l=0}^{M_j}$, given $\{x_i^{(t-1)}\}_{i=1}^n$, $\{s_{i1}^{(t-1)},\ldots,s_{iJ^E}^{(t-1)}\}_{i=1}^n$, and $\{\alpha_j^{(t-1)},\beta_j^{(t-1)}\}_{j=1}^{J^E}$ using the Metropolis-Hastings algorithm of Cowles (1996). Throughout the iterations, we fix $\tau_{0j}^{(t)}=-\infty$, $\tau_{1j}^{(t)}=0$, and $\tau_{M_j,j}^{(t)}=\infty$ for each $j$.

**Step 2:** Sample $(\alpha_j^{(t)},\beta_j^{(t)})$ given $\{x_i^{(t-1)}\}_{i=1}^n$, $\{s_{i1}^{(t-1)},\ldots,s_{iJ^E}^{(t-1)}\}_{i=1}^n$, $\{\tau_{lj}^{(t)}\}_{l=0}^{M_j}$ for each $j=1,\ldots,J^E$.

This can be accomplished in the following two steps.

$$\begin{aligned}
\widetilde{Y}_{ij}^{E^{(t)}} \mid x_i^{(t-1)}, s_{ij}^{(t-1)}, \alpha_j^{(t-1)}, \beta_j^{(t-1)} &\sim \mathbf{1}\left\{\tau_{Y_{ij}^E,j}^{(t)} \leq \widetilde{Y}_{ij}^{E^{(t)}} \leq \tau_{Y_{ij}^E+1,j}^{(t)}\right\} \mathcal{N}\left(-\alpha_j^{(t-1)} + \beta_j^{(t-1)}\left(x_i^{(t-1)} + s_{ij}^{(t-1)}\right),\ 1\right) \\
(\alpha_j^{(t)}, \beta_j^{(t)}) \mid \{\widetilde{Y}_{ij}^{E^{(t)}}, x_i^{(t-1)}, s_{ij}^{(t-1)}\}_{i=1}^n &\sim \mathbf{1}\{\beta_j^{(t)} \geq 0\}\ \mathcal{N}\left(\Lambda^{(t-1)} \sum_{i=1}^n W_i^{(t-1)} \widetilde{Y}_{ij}^{E^{(t)}},\ \Lambda^{(t-1)}\right)
\end{aligned}$$

where $\Lambda^{(t-1)} = \left(\sum_{i=1}^n W_i^{(t-1)} W_i^{(t-1)^\top} + B\right)^{-1}$ and $W_i = \left(-1\ , x_i^{(t-1)} + s_{ij}^{(t-1)}\right)^\top$. The second step is implemented by first drawing $\alpha_j^{(t)}$ from its marginal distribution, which is a univariate normal distribution, and then sampling $\beta_j^{(t)}$ from its conditional distribution given $\alpha_j^{(t)}$, which is a univariate truncated normal distribution.

**Step 3:** Sample $s_{ij}^{(t)}$ given $x_i^{(t-1)}$, $\widetilde{Y}_{ij}^{E^{(t)}}$, $\alpha_j^{(t)}$, $\beta_j^{(t)}$, $\lambda^{(t-1)}$, and $\omega^{2(t-1)}$ for each $(i, j)$. This step can be accomplished by the standard Gibbs sampling algorithm of the Bayesian normal regression for a single observation where the outcome variable is $\widetilde{Y}_{ij}^{E^{(t)}} + \alpha_j^{(t)} - \beta_j^{(t)} x_i^{(t-1)}$, the predictor is $\beta_j^{(t)}$, the error variance is fixed at 1, and the prior distribution for $s_{ij}^{(t)}$ is $\mathcal{N}(V_i^\top \lambda^{(t-1)},\ \{\omega^{(t-1)}\}^2)$.

**Step 4:** Sample $x_i^{(t)}$ given $\{s_{ij}^{(t)}, \widetilde{Y}_{ij}^{E^{(t)}}, \alpha_j^{(t)}, \beta_j^{(t)}\}_{j=1}^{J^E}$, and $\delta^{(t-1)}$ for each $i$. This step can be accomplished by the standard Gibbs sampling algorithm of the Bayesian normal regression where the outcome variable is $\widetilde{Y}_{ij}^{E^{(t)}} + \alpha_j^{(t)} - \beta_j^{(t)} s_{ij}^{(t)}$, the predictor is $\beta_j^{(t)}$, the error variance is fixed at 1, and the prior distribution for $x_i$ is $\mathcal{N}(V_i^\top \delta^{(t-1)}, 1)$.

**Step 5:** Sample $\lambda^{(t)}$ and $\omega^{(t)}$ given all $s_{ij}^{(t)}$. This step can be accomplished by the standard Gibbs sampling algorithm of the Bayesian normal regression where the outcome variable is $s_{ij}^{(t)}$, the predictor is $V_i$, the vector of coefficients is $\lambda^{(t)}$, the error variance is $\{\omega^{(t)}\}^2$, and the prior distribution for $\lambda^{(t)}$ is $\mathcal{N}(0, D)$.

**Step 6:** Sample $\delta^{(t)}$ given $x_i^{(t)}$. This step can be accomplished by the standard Gibbs sampling algorithm of the Bayesian normal regression where the outcome is $x_i^{(t)}$, the predictor is $V_i$, and the error variance is fixed at 1.

## A.3 The Combined Model

To combine the two models, we assume $\gamma = \lambda/\omega$. All the steps of the MCMC algorithm are identical to those described in Appendices A.1 and A.2 except that Step 2 of Appendix A.1 and Step 4 of Appendix A.2 will be combined into the standard updating of a stacked regression model where the dependent variable consists of $s_{ij}^{(t)}$ as well as $Z_i^{*(t)}$ (which is now sampled from the truncated normal with mean $V_i^\top \lambda^{(t-1)}$ and standard deviation $\omega^{(t-1)}$), the independent variable is $V_i$, and the variance parameter is $\{\omega^{(t-1)}\}^2$. We can use the standard updating procedure with the semi-conjugate prior distribution.

## A.4 Estimated Coefficients for the Three Models

| | List Experiment | | Endorsement Experiment | | Combined | |
|---|---|---|---|---|---|---|
| | est. | s.e. | est. | s.e. | est. | s.e. |
| *Individual-level* | | | | | | |
| Harm from Taliban violence | 0.41 | 1.29 | 0.45 | 0.32 | 0.54 | 0.23 |
| Harm from Taliban violence is NA | 2.23 | 2.52 | −0.67 | 0.80 | −0.54 | 0.63 |
| Harm from ISAF violence | −2.69 | 1.48 | −0.36 | 0.25 | −0.34 | 0.17 |
| Harm from ISAF violence is NA | 0.99 | 2.75 | 2.54 | 1.19 | 1.66 | 0.81 |
| Approach by Taliban after Harm | 1.83 | 1.45 | −1.28 | 0.42 | −1.00 | 0.30 |
| Approach by Taliban after Harm is NA | 0.24 | 3.19 | 1.59 | 1.96 | 2.06 | 1.44 |
| Approach by ISAF after Harm | 3.77 | 1.70 | 0.42 | 0.47 | 0.45 | 0.34 |
| Approach by ISAF after Harm is NA | 0.21 | 2.94 | 2.51 | 2.30 | 2.38 | 2.00 |
| ISAF encounter frequency | 1.08 | 0.43 | 0.10 | 0.13 | 0.17 | 0.09 |
| Years of education | 0.12 | 0.08 | 0.02 | 0.03 | 0.01 | 0.02 |
| Age (tens) | 0.46 | 0.29 | −0.09 | 0.09 | 0.02 | 0.06 |
| Income (Afghanis) | 0.09 | 0.46 | 0.33 | 0.16 | 0.17 | 0.10 |
| Income is NA | −0.47 | 1.38 | 0.26 | 0.60 | 0.01 | 0.41 |
| Schooled in madrassa | −1.62 | 0.67 | −0.44 | 0.22 | −0.22 | 0.15 |
| Pro-Taliban tribe | −1.99 | 1.00 | 0.05 | 0.29 | 0.07 | 0.21 |
| Pro-Taliban tribe is NA | 0.69 | 1.41 | −0.13 | 0.48 | −0.22 | 0.32 |
| *Village-level* | | | | | | |
| Altitude (km) | −0.87 | 0.48 | −0.16 | 0.14 | −0.07 | 0.10 |
| Population | −0.21 | 0.68 | 0.09 | 0.11 | 0.07 | 0.08 |
| ISAF-initiated violent events (within 5km) | 1.14 | 0.82 | −0.03 | 0.14 | 0.01 | 0.10 |
| Taliban-initiated violent events (within 5km) | −2.72 | 1.14 | −0.07 | 0.14 | −0.10 | 0.12 |
| *District-level* | | | | | | |
| Sha'ria courts | −2.01 | 2.24 | −0.56 | 0.40 | −0.04 | 0.28 |
| CERP project spending | 3.69 | 1.39 | −0.01 | 0.23 | −0.08 | 0.16 |
| Opium cultivation (ha.) | −6.24 | 1.37 | 0.08 | 0.19 | −0.09 | 0.14 |
| CDC project count | −0.81 | 0.52 | −0.16 | 0.11 | −0.00 | 0.08 |
| Road length (km) | 1.93 | 0.72 | 0.17 | 0.15 | 0.19 | 0.11 |
| Pakistan border | 0.69 | 1.35 | 0.26 | 0.37 | 0.07 | 0.25 |
| Government territorial control | −4.52 | 2.01 | 0.43 | 0.41 | 0.23 | 0.30 |
| Contested territorial control | −2.67 | 1.31 | −0.33 | 0.28 | 0.00 | 0.19 |
| *Intercept* | −5.01 | 1.93 | −0.82 | 0.60 | −1.58 | 0.50 |

## A.5 Sensitivity Analysis

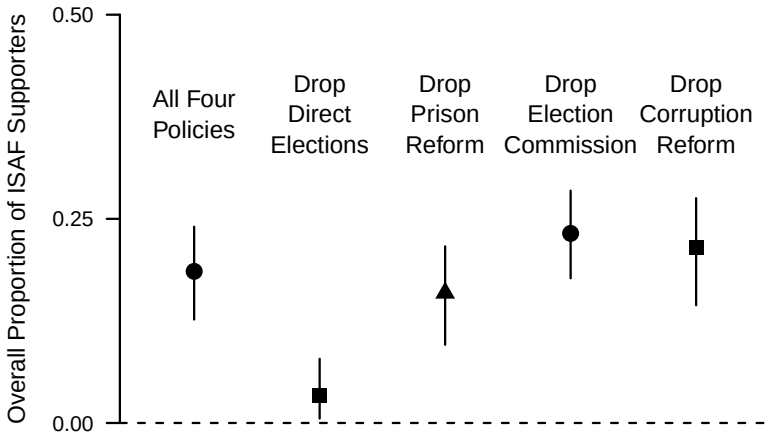


Figure 8: Estimated Mean Support Levels, Dropping One of the Four Policy Questions. This figure presents the estimates of the same quantities of interest as those in Figure 5 based on the models where one of the four policies is excluded from the analysis.

| | All Four Policies | | Drop Direct Elections | | Drop Prison Reform | | Drop Election Commission | | Drop Corruption Reform | |
|---|---|---|---|---|---|---|---|---|---|---|
| | est. | s.e. | est. | s.e. | est. | s.e. | est. | s.e. | est. | s.e. |
| *Individual-level* | | | | | | | | | | |
| Harm from Taliban violence | 0.45 | 0.32 | −0.51 | 1.18 | 0.61 | 0.35 | 0.40 | 0.26 | 0.49 | 0.36 |
| Harm from Taliban violence is NA | −0.67 | 0.80 | −1.04 | 1.58 | −0.52 | 0.89 | −0.36 | 0.65 | −1.22 | 0.91 |
| Harm from ISAF violence | −0.36 | 0.25 | −0.88 | 1.06 | −0.46 | 0.27 | −0.43 | 0.20 | −0.36 | 0.28 |
| Harm from ISAF violence is NA | 2.54 | 1.19 | 0.69 | 1.82 | 2.66 | 1.43 | 1.85 | 0.97 | 3.35 | 1.43 |
| Approach by Taliban after Harm | −1.28 | 0.42 | −3.41 | 2.30 | −1.68 | 0.55 | −0.84 | 0.33 | −1.08 | 0.47 |
| Approach by Taliban after Harm is NA | 1.59 | 1.96 | 0.13 | 1.78 | 1.32 | 2.18 | 1.11 | 1.52 | 1.83 | 2.17 |
| Approach by ISAF after Harm | 0.42 | 0.47 | 1.69 | 1.78 | 0.38 | 0.53 | 0.50 | 0.39 | 0.08 | 0.53 |
| Approach by ISAF after Harm is NA | 2.51 | 2.30 | 0.11 | 1.79 | 2.55 | 2.54 | 1.74 | 1.85 | 3.46 | 2.62 |
| ISAF encounter frequency | 0.10 | 0.13 | 0.33 | 0.71 | 0.09 | 0.14 | 0.04 | 0.11 | 0.11 | 0.14 |
| Years of education | 0.02 | 0.03 | 0.10 | 0.11 | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 |
| Age (tens) | −0.09 | 0.09 | −0.77 | 0.60 | −0.05 | 0.10 | −0.08 | 0.07 | −0.12 | 0.10 |
| Income (Afghanis) | 0.33 | 0.16 | 0.89 | 0.77 | 0.25 | 0.17 | 0.29 | 0.13 | 0.37 | 0.18 |
| Income is NA | 0.26 | 0.60 | −0.51 | 1.43 | 0.16 | 0.67 | 0.38 | 0.49 | 0.26 | 0.67 |
| Schooled in madrassa | −0.44 | 0.22 | −2.05 | 1.47 | −0.33 | 0.24 | −0.41 | 0.18 | −0.52 | 0.26 |
| Pro-Taliban tribe | 0.05 | 0.29 | 0.04 | 1.02 | −0.01 | 0.33 | 0.10 | 0.24 | 0.14 | 0.33 |
| Pro-Taliban tribe is NA | −0.13 | 0.48 | −0.08 | 1.36 | −0.23 | 0.53 | −0.05 | 0.40 | −0.20 | 0.54 |
| *Village-level* | | | | | | | | | | |
| Altitude (km) | −0.16 | 0.14 | −0.70 | 0.79 | −0.20 | 0.17 | −0.18 | 0.12 | −0.15 | 0.16 |
| Population | 0.09 | 0.11 | 0.56 | 0.64 | 0.07 | 0.12 | 0.09 | 0.09 | 0.08 | 0.12 |
| ISAF-initiated violent events (within 5km) | −0.03 | 0.14 | 0.47 | 0.67 | 0.02 | 0.15 | −0.08 | 0.12 | −0.07 | 0.15 |
| Taliban-initiated violent events (within 5km) | −0.07 | 0.14 | −0.39 | 0.67 | −0.07 | 0.15 | −0.02 | 0.11 | −0.12 | 0.15 |
| *District-level* | | | | | | | | | | |
| Sha'ria courts | −0.56 | 0.40 | −0.98 | 1.46 | −0.72 | 0.45 | −0.49 | 0.33 | −0.76 | 0.48 |
| CERP project spending | −0.01 | 0.23 | 0.14 | 0.87 | −0.01 | 0.25 | −0.03 | 0.19 | −0.06 | 0.26 |
| Opium cultivation (ha.) | 0.08 | 0.19 | −0.80 | 0.85 | 0.05 | 0.20 | 0.09 | 0.15 | 0.23 | 0.22 |
| CDC project count | −0.16 | 0.11 | −0.58 | 0.55 | −0.19 | 0.13 | −0.12 | 0.09 | −0.23 | 0.13 |
| Road length (km) | 0.17 | 0.15 | 0.72 | 0.81 | 0.19 | 0.17 | 0.10 | 0.12 | 0.22 | 0.18 |
| Pakistan border | 0.26 | 0.37 | 0.64 | 1.35 | 0.28 | 0.41 | 0.18 | 0.31 | 0.34 | 0.42 |
| Government territorial control | 0.43 | 0.41 | 1.26 | 1.37 | 0.38 | 0.46 | 0.37 | 0.34 | 0.51 | 0.47 |
| Contested territorial control | −0.33 | 0.28 | −0.51 | 1.02 | −0.21 | 0.29 | −0.33 | 0.22 | −0.37 | 0.32 |
| *Intercept* | −0.82 | 0.60 | −2.85 | 2.16 | −0.84 | 0.66 | −0.54 | 0.48 | −0.49 | 0.67 |

Table 2: Comparison of the Estimated Coefficients for the Endorsement Experiment Models that Include All Four Questions and the Models that Exclude One of the Four Policy Questions.

# References


Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2011. "Winning Hearts and Minds? Evidence from a Field Experiment in Afghanistan.". MIT Political Science Working Paper No.2011-14.

Berman, Eli, Jacob Shapiro, and Joseph Felter. 2011. "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq." *Journal of Political Economy* 119: 766-819.

Blair, Graeme, Christine Fair, Neil Malhotra, and Jacob Shapiro. 2013. "Poverty and Support for Militant Politics: Evidence from Pakistan." *American Journal of Political Science* 57 (1): 30–48.

Blair, Graeme, and Kosuke Imai. 2011. "list: Statistical Methods for the Item Count Technique and List Experiment." available at the Comprehensive R Archive Network (CRAN). `http://CRAN.R-project.org/package=list`.

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

Bullock, Will, Kosuke Imai, and Jacob N. Shapiro. 2011. "Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan." *Political Analysis* 19 (Autumn): 363–384.

Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers?: Modeling the List Experiment with LISTIT." *Political Analysis* 17 (1): 45–63.

Cowles, Mary Kathryn. 1996. "Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models." *Statistics and Computing* 6 (2): 101–111.

Fisher, R. A. 1915. "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Independently Large Population." *Biometrika* 10 (May): 507–521.

Flavin, Patrick, and Michael Keane. 2010. How Angry am I? Let Me Count the Ways: Question Format Bias in List Experiments. Technical report Department of Political Science, University of Notre Dame.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15 (Winter): 1–20.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulations Using Multiple Sequences (with Discussion)." *Statistical Science* 7 (4): 457–472.

Gingerich, Daniel W. 2010. "Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys." *Political Analysis* 18 (Summer): 349–380.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77: 159-172.

Gonzalez-Ocantos, Ezequiel, Chad Kiewet de Jonge, Carlos Melendez, Javier Osorio, and David W. Nickerson. 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56 (1): 202-217.

Hawkins, D. L. 1989. "Using U Statistics to Derive the Asymptotic Distribution of Fisher's Z Statistic." *The American Statistician* 43 (November): 235–237.

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 (Spring): 37–67.

Imai, Kosuke. 2011. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association* 106 (June): 407–416.

Kalyvas, Stathis. 2006. *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the "New South"." *Journal of Politics* 59 (May): 323–349.

Leites, Nathan, and Charles Wolf. 1970. *Rebellion and Authority: An Analytic Essay on Insurgent Conflicts*. Chicago: Markham Publishing Company.

Lyall, Jason, Graeme Blair, and Kosuke Imai. 2013. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." *American Political Science Review*. Forthcoming.

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

Raghavarao, D., and W. T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society, Series B, Methodological* 41 (1): 40–45.

Transue, John E., Daniel J. Lee, and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17 (Spring): 143–161.

Tsuchiya, Takahiro, Yoko Hirai, and Shigeru Ono. 2007. "A Study of the Properties of the Item Count Technique." *Public Opinon Quarterly* 71 (Summer): 253–272.

U.S. Army. 2007. *U.S. Army Field Manual No. 3-24*. Chicago: University of Chicago Press.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60 (March): 63–69.