Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan*

Will Bullock† Kosuke Imai‡ Jacob N. Shapiro§

First Draft: June 7, 2010
This Draft: March 11, 2011

Abstract

Political scientists have long been interested in citizens' support level for socially sensitive actors such as ethnic minorities, militant groups, and authoritarian regimes. Attempts to use direct questioning in surveys, however, have largely yielded unreliable measures of these attitudes as they are contaminated by social desirability bias and high non-response rates. In this paper, we develop a statistical methodology to analyze endorsement experiments, which recently have been proposed as a possible solution to this measurement problem. The commonly used statistical methods are problematic because they cannot properly combine responses across multiple policy questions, the design feature of a typical endorsement experiment. We overcome this limitation by using item response theory to estimate support levels on the same scale as the ideal points of respondents. We also show how to extend our model to incorporate a hierarchical structure of data in order to recoup the loss of statistical efficiency due to indirect questioning. We illustrate the proposed methodology by applying it to measure political support for Islamist militant groups in Pakistan. Simulation studies suggest that the proposed Bayesian model yields estimates with reasonable levels of bias and statistical power. Finally, we offer several practical suggestions for improving the design and analysis of endorsement experiments.

*Financial support from the NSF grant (SES–0849715 for Imai), DHS (2007-ST-061-000001 for Shapiro) through the National Center for Risk and Economic Analysis of Terrorism Events, and the Mamdouha S. Bobst Center for Peace and Justice (for Imai and Shapiro) is acknowledged. The Institute for Quantitative Social Science at Harvard University provided the computational support necessary for simulation studies. We thank John Londregan, Nolan McCarty, Michael Peress, and seminar participants at Columbia and Princeton Universities for helpful comments. Christine Fair and Neil Malhotra helped design the survey, to which we apply the measurement model presented in this paper. We also thank Joshua Borkowski for his help in creating the map presented in Figure 4. All of the code necessary to reproduce the results in this paper can be downloaded from the Dataverse as **?**. A previous version of this paper was circulated as "Measuring Political Support and Issue Ownership Using Endorsement Experiments, with Application to the Militant Groups in Pakistan."

†Graduate Student, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–0092, E-mail: wbullock@princeton.edu, URL: http://wbullock.princeton.edu

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

§Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–2256, Email: jns@princeton.edu, URL: http://www.princeton.edu/∼jns/

1 Introduction

Political scientists have long been interested in measuring sensitive political attitudes. Yet, asking sensitive questions directly in surveys has often resulted in bias due to social desirability and high item non-response. To overcome these problems, researchers have relied upon survey methodologies that are specifically designed for eliciting truthful responses through the means of indirect questioning. Two most commonly used techniques are randomized response methods and list experiments (see e.g., **??**).

In this paper, we focus on an alternative survey methodology called *endorsement experiments*, which have been used to measure the levels of political support for socially sensitive actors such as ethnic minorities, militant groups, and authoritarian regimes.[1] In endorsement experiments, randomly selected respondents are asked to express their opinion about several policies endorsed by a socially sensitive actor of interest. These responses are then contrasted with those from a control group that receives no endorsement. If the endorsement by a political actor induces more support for policies, then this is taken as evidence for the existence of support for that actor. The main advantage of endorsement experiments is that indirect questioning may increase truthful responses and improve response rate. The drawback, on the other hand, is that it can only provide an indirect measure of support.

Despite the increasing popularity and promise of endorsement experiments, the common statistical methods used to analyze endorsement experiments (e.g., difference in means and regression models) have been unsatisfactory. In particular, the standard methods fail to properly combine responses across multiple policy questions, the design feature of a typical endorsement experiment. To obtain robust estimates of support level, most researchers use multiple policy questions. Yet, there exists no principled way of combining responses to these different questions to produce a single measure of support level. Typical endorsement experiments also attempt to measure support levels for multiple groups. This poses another statistical challenge because one must deal with relatively small sample size per group. Together with the loss of information due to indirect questioning, this calls for a statistical method that allows researchers to efficiently estimate levels of support for multiple group by combining responses to several policy questions.

To overcome these methodological challenges, we develop a Bayesian measurement model to analyze such endorsement experiments.[2] Our model is based on the item response theory, which has been used in political science to estimate ideal points of legislators from roll call data (e.g.,

[1] See **?**, **?**, and **?** for recent applications in political science. **?** also use a similar experiment with endorsements by Congress and community organizations.

[2] All of the code necessary to reproduce the results in this paper can be downloaded from the Dataverse as **?**.

???).[3] The proposed model provides estimates of political support measured on the same scale as the ideal points of respondents by combining responses to multiple policy questions. This framework facilitates a meaningful interpretation of the magnitude of support measures.[4]

Furthermore, our model can incorporate individual level covariates to explore the association between respondents' characteristics and the degree of their support for a political actor. Thus, the proposed methodology can efficiently recover the latent support level while directly incorporating all the estimation uncertainty.[5] We also show that our model can efficiently accommodate a multilevel structure of data via hierarchical modeling (**?**). This means that the proposed model can incorporate aggregate level (e.g., geographical units) covariates as well as individual characteristics. Finally, we show how to conduct post-stratification to further obtain efficient aggregate level estimates of support based on census data whenever such information is available (**??**).

The proposed methodology clarifies the key assumptions that make it possible for researchers to interpret the results of endorsement experiments as measures of support levels. This in turn enables us to offer practical suggestions for improving the design and analysis of endorsement experiments so that the credibility of these assumptions can be increased. These suggestions are summarized in the concluding section of this paper.

We offer both empirical and simulation evidence demonstrating the applicability and statistical efficiency of the proposed methodology. On the empirical side, we apply our model to a recent large scale endorsement experiment in Pakistan and estimate levels of support among citizens for different militant groups such as al-Qa'ida. We show that the proposed method yields efficient estimates of support for each group within divisions, which are relatively small geographical units of Pakistan. The efficiency gains are especially valuable in this context as they allow us to study spatial variation in support for militant organizations across more refined political boundaries than is possible with standard methods.

Our analysis reveals interesting patterns of support for militant groups, which are not obtained by standard regression models. First, attitudes towards militant groups are quite geographically clustered. Even within the four provinces of Pakistan there are clear clusters of support and enmity towards these groups, with Gujranwala in northern Punjab emerging as the most supportive region for most groups. This is in marked contrast to a policy discourse that focuses on Bahawalpur, in

[3] Although some have used the item response theory to analyze surveys (e.g., **??**), to the best of our knowledge none have applied it to survey experiments.

[4] The model can also yield an estimated level of support for each political actor given any particular policy rather than averaging across all policies.

[5] This contrasts with the standard two-step estimation procedure where the levels of support are first estimated and then these estimates are regressed on individual covariates of interest.

southern Punjab, as the locus of support for militancy in Pakistan. Second, militant groups tend to receive less support in the areas where they conduct the most attacks. In particular, Pakistanis' support for the Afghan Taliban is almost inversely proportional to their distance from the Afghan border. This finding raises a subtle point, that high levels of militant activity in an area should not be taken as evidence the population there supports militant groups. Third, we find that the individual level variables that are the focus of much policy attention — education and income, for example, are frequently cited as key correlates of support for Islamist militancy — matter little once regional differences are taken into account. It is thus local political attitudes that are the key determinate of support, a finding with strong implications for policies to combat the threat of militancy in Pakistan.

On the simulation side, we show that our Bayesian measurement model has good (frequentist) statistical properties. Our first set of simulation studies are based on the data generating process that closely follows the Pakistan data which have a relatively large sample size. In this setup, we show that the proposed model can recover the true parameter values with little bias and strong statistical power. In addition, our Bayesian confidence intervals have the coverage rate close to the nominal rate. The second set of simulation studies uses a similar setup but is based on sample sizes of typical surveys. Here, we find that the bias of the resulting estimates remains small and that the statistical power decreases, yet stays at a reasonable level.

The rest of the paper is organized as follows. In Section 2, we briefly describe the recent large scale endorsement experiment conducted in Pakistan, to which we apply our proposed methodology. In Section 3, we introduce our proposed Bayesian hierarchical measurement model for endorsement experiments and formally define quantities of interest. We also discuss issues related to estimation and inference based on this model. In Section 4, we apply our method to the Pakistan data and present the empirical results. In Section 5, we show the results of simulation studies and evaluate the statistical properties of the proposed Bayesian measurement model. Finally, in Section 6, we offer concluding remarks and practical suggestions for improving the design and analysis of endorsement experiments.

2 Measuring the Support for Militant Groups in Pakistan

2.1 Background

Militant violence in Pakistan stands at the top of the international security agenda, yet little is known about who supports militant organizations and why. As a result, discussions about why Pakistan suffers so much political violence tend to turn to untested assertions that poverty, poor education, and resistance to Western values drive support for militant organizations (see **?**, for a summary of

these arguments). It is hard to overstate the influence these ideas have. U.S. and Western policies towards Pakistan have devoted billions of dollars to encouraging economic and social development as an explicit means of diminishing the militant threat.

In some ways, this state of affairs is not surprising. Directly asking respondents how they feel about militant organizations is problematic, to say the least, in places suffering from political violence. Beyond safety concerns for enumerators and respondents who discuss such issues, item non-response rates are often quite high as respondents understandably fear that providing the "wrong" answer will lead to social censure and may even threaten their safety. Moreover, direct questions are subject to social desirability bias in so far as answers combine respondents' true attitudes with what they believe to be the socially appropriate response.

One way to overcome these problems is through the use of an endorsement experiment. This approach measures the differences in support for various policies unrelated to militancy between two groups; one group of respondents who are told only about the policy and another group who are told that a militant organization endorses it. If respondents are randomly divided in this way, then the differences between the two conditions reveals how much support for a policy changes by being associated with a militant group, thereby providing an indirect measure of support for that group. Unlike a direct measure, non-response and social desirability biases are minimized since respondents are reacting to the policy and not to the group itself. By asking respondents about multiple policy issues and randomizing the pairing of issue with group, we can identify effects for multiple groups that are unlikely to be influenced by the details of any specific policy.

2.2 The Pakistan Survey Experiment

? implement such a design to study support for four groups — Pakistani militants fighting over Kashmir, militants fighting in Afghanistan (a.k.a. the Afghan Taliban), al-Qa'ida, and the sectarian militias or *firqavarana tanzeems*—using a large, nationally-representative sample of Pakistanis. The survey entailed two main innovations. First, as noted, the researchers used an endorsement experiment to avoid the high item non-response rates and social desirability bias that plagued previous efforts to study the politics of militancy in Pakistan. Second, the researchers employed a sample that permitted them to make reliable inferences about a range of political attitudes in each of the four provinces of Pakistan: Punjab, Sindh, Khyber-Pakhtunwa (KPK), and Balochistan.

Working with Pakistani partners, Socio-Economic Development Consultants (SEDCO), the researchers drew a sample of 6,000 adult Pakistani men and women. The respondents were selected randomly within 500 primary sampling units (PSU). Following the rural/urban breakdown in the Pakistan census, the survey entailed 332 rural PSU and 168 urban ones. Because Pakistan's provinces

are heavily skewed — the KPK and Balochistan only account for 14.5% and 4.9% of Pakistan's population respectively (**?**, p.68) — the survey over-sampled these smaller provinces. The face-to-face questionnaire was fielded by six mixed-gender teams between April 21, 2009 and May 25, 2009.

Perhaps in part due to the indirect questioning, the overall response rate was over 90 percent, approaching the response rates achieved by the United States Census Bureau. The Pakistan endorsement experiment had the following features. First, respondents were randomly assigned to treatment or control groups (one-half of the sample was assigned to each group). Respondents in the control group were asked their level of support for four policies:

- World Health Organization (WHO) plan to provide universal polio vaccinations in Pakistan.
- Government of Pakistan plan for curriculum reforms in religious schools or *madaris*.
- Reforming the Frontier Crimes Regulation (FCR) governing the tribal areas.
- Using peace jirgas to resolve disputes over the Afghan border, the Durand Line.

Respondents in the treatment group were asked identical questions but were told that one of four groups supports the policy in question. The group associated with each of the four policies was randomized within the treatment group, effectively randomizing the order of the endorsements.

For example, the script for the control group used in this survey experiment reads as follows for the WHO polio vaccination plan question.

```
The World Health Organizations recently announced a plan to introduce universal
Polio vaccination across Pakistan.  How much do you support such a plan?

(1) A great deal;  (2) A lot;  (3) A moderate amount; (4) A little; (5) Not at all
```

In contrast, the script for a treatment group with respect to the same policy question reads as follows,

```
The World Health Organizations recently announced a plan to introduce universal
Polio vaccination across Pakistan.  Pakistani militant groups fighting in Kashmir
have voiced support for this program.  How much do you support such a plan?

(1) A great deal;  (2) A lot;  (3) A moderate amount; (4) A little; (5) Not at all
```

These four policies were chosen in consultation with SEDCO and other Pakistani colleagues to meet three criteria. First, each policy issue was being discussed in Pakistan at the time. Second, an informed citizen would be likely to know about these policies. Third, citizens' opinions over the policies were not rigid so that the group endorsements could have an effect. All three criteria were met based on the results of a 200 person pre-test and four focus groups conducted during the

survey-design phase (see also Section 6 for practical suggestions about how to improve the design of endorsement experiments).

Figure 1 graphically displays the aggregate distribution of responses across four provinces from the Pakistan survey experiment. Pakistanis held generally favorable views towards each of the four policies, with greatest support for education reforms and polio vaccinations. The plots show that there exist significant differences across provinces (four rows of the graph) for each policy question. On the other hand, in each province, the aggregate distribution of responses does not vary dramatically between the control group and each treatment group for any of the four policy questions. Thus, the main methodological challenge is whether we will be able to detect any systematic pattern of support levels from this data using the proposed method.

2.3 Substantive Issues

Before turning to a detailed discussion of the proposed method, we briefly highlight key substantive issues we hope to address in our empirical analysis. First, as with all political organizations, militant groups have highly localized components. Understanding the patterns of support in more detail can help identify the specific grievances that drive people to embrace these organizations. Second, in addition to the spatial variation of support levels for militant groups, researchers are interested in how individual characteristics – education, personal religiosity, and socio-economic status, for example – are associated with support levels after taking into account regional differences.

Finally, we also investigate the relationship between levels of political violence and support for militant groups. There is striking variation in levels of violence across Pakistan. In the year before the survey was fielded, for example, the number of people killed in terrorist attacks ranged from 11,429 in Khyber-Pakhtunwa province (population of 17 million), to 1,791 in Sindh province (population of 30 million), almost an 10-fold difference in per capita violence. And within each province, there exists substantial further variation in violence across smaller political units. Understanding how levels of violence correlate with support for militant groups can provide vital insight into how militants interact with the population at large. If support is higher in more violent areas, that could suggest that popular approval is a key enabler for violent groups. If, however, support is lower in more violent areas, that could suggest that the negative externalities of terrorist violence may actually alienate the population, at least locally.

Answering all of these substantive questions requires researchers to obtain precise estimates of support within relatively small geographical boundaries. Standard methods of analyzing endorsement experiments cannot do so without prohibitively large samples. We therefore turn to the proposed methodology, which allows us to combine responses to multiple policy questions in a principled

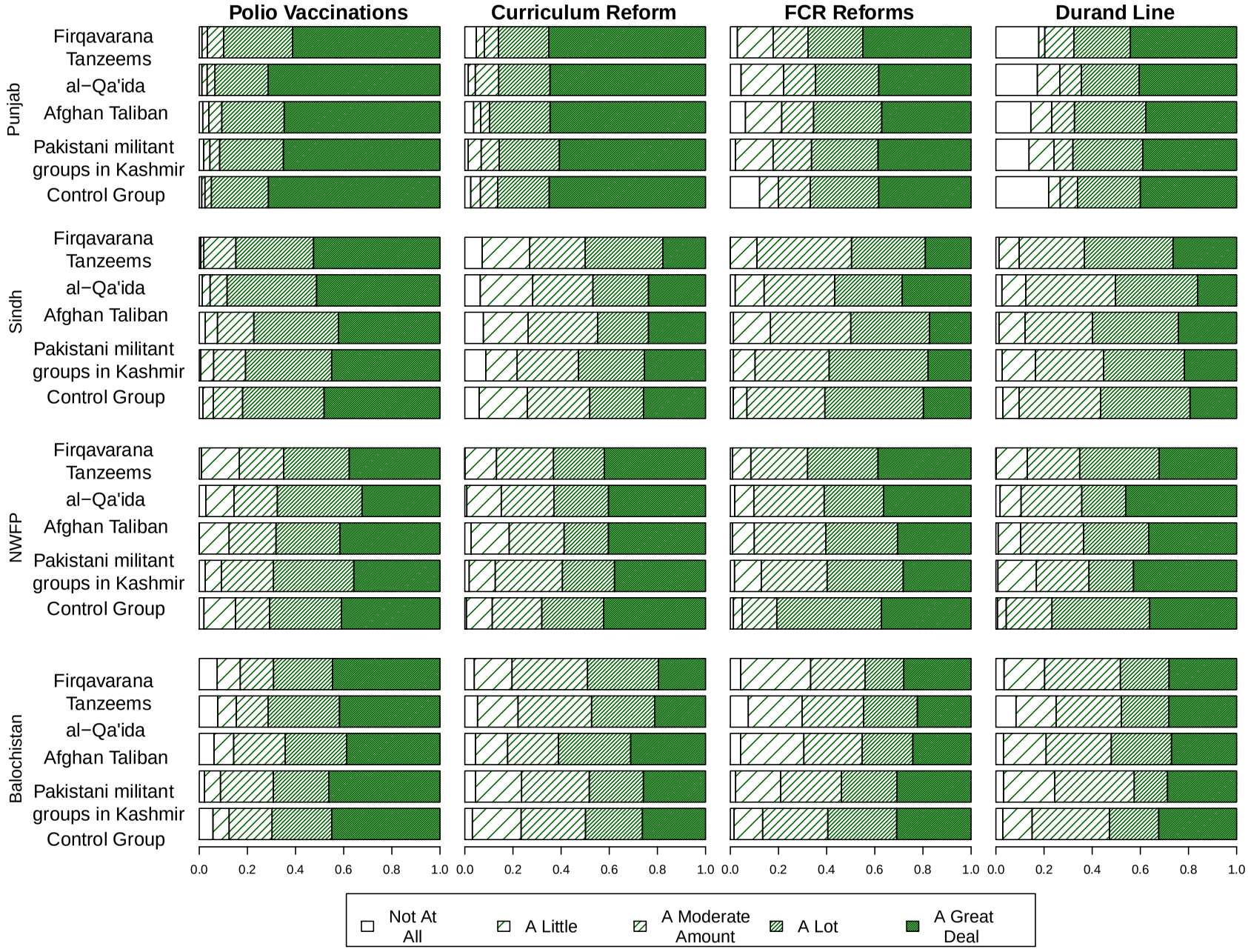

Figure 1: Aggregate Distribution of Responses from the Pakistan Survey Experiment. The aggregate distribution of responses varies most notably across issues and provinces, but does not vary dramatically between the control group and each treatment group for any of the four policy questions.

manner and obtain efficient estimates of support levels for several militant groups at both individual and aggregate levels.

3 The Proposed Methodology

In this section, we introduce a new methodology for analyzing endorsement experiments. We begin by formally describing the design of the survey experiment, and then introduce the Bayesian hierarchical measurement model of political support. Under this model, we define quantities of interest and briefly explain the issues related to estimation and inference.

3.1 The Basic Design of Endorsement Experiments

Suppose that we wish to measure the level of support for K political actors within a target population $\mathcal{P}$. To do so, we consider a survey experiment for a random sample of N respondents from $\mathcal{P}$. In the survey, respondents are asked whether or not they support each of J policies chosen by researchers. Here, for the sake of simplicity, we begin by assuming that respondents are answering a binary question regarding their support for a policy. We later generalize our model to ordinal responses so that respondents can be asked to specify the degree of support for a policy using an ordinal scale, as done in the Pakistani experiment. We assume that these J policies are a representative sample of a population pool of policies $\mathcal{J}$ (i.e., a policy space), with respect to which the level of support for political actors is to be measured. If we index each policy by j, we can use Y_{ij} to represent respondent i's answer to the survey question regarding policy j where $Y_{ij} = 1$ ($Y_{ij} = 0$) implies respondent i favors (opposes) policy j. We denote an M dimensional vector of the observed characteristics of respondent i by Z_i.

The key aspect of the survey experiment is to attach the endorsement by a particular political actor to each of these policy questions. We then observe whether the level of support for these policies changes in comparison to the case where no such endorsement is given. We randomly assign one of K political actors as an endorser to respondent i's question regarding policy j and denote this "treatment" variable by $T_{ij} \in \{0, 1, \ldots, K\}$. We use $T_{ij} = 0$ to represent the "control" observations where no political endorsement is attached to the question. Alternatively, one may use the endorsement by a neutral actor as the control group.

Using the standard statistical framework of causal inference (**?**), this implies that we can define $(K + 1)$ potential outcomes for respondent i's answer to the question regarding policy j. Thus, $Y_{ij}(0)$ represents respondent i's potential answer without any political endorsement and $Y_{ij}(k)$ is respondent i's potential answer when policy j is endorsed by political actor k. Although we only observe one of these potential outcomes, i.e., $Y_{ij} = Y_{ij}(T_{ij})$, the randomization of the treatment

implies that the average treatment effect can be identified using the difference in means between treatment and control groups because of the independence between the treatment and the potential outcomes.[6]

There are a number of challenges that researchers must overcome in order to effectively use endorsement experiments. First and foremost, researchers must choose a set of policies to ask about and our analysis provides some guidance here. Other choices include question wording, question ordering, and possible carryover effects from one question to another. While the discussion of these other design issues is beyond the scope of this paper, we emphasize that researchers must take these issues seriously before applying the proposed statistical methodology. We provide some practical suggestions about how to design endorsement experiments in Section 6.

3.2 The Bayesian Measurement Model

Given the basic design of endorsement experiments described above, we propose a Bayesian measurement model of political support. The main feature of the proposed model is to analyze survey experiments within the framework of the standard ideal point models used in political science.

The Basic Binary Response Model. We begin by assuming that the quadratic utility function can completely characterize each respondent's answer to the question regarding a particular policy. Following the notation of **?**, let ζ_{j1} and ζ_{j0} denote the location of favoring and opposing policy j in the single dimension policy space.[7] Note that it is also possible to generalize this model to multi-dimensional policy space, but such an extension is likely to be of little practical use given the limited number of issues typically used in endorsement experiments. As such, we recommend that researchers who want to use this framework carefully choose policy questions from a single policy dimension when designing endorsement experiments.

Given this setup, using x_i to denote respondent i's ideal point, we can write the quadratic utility functions as,

$$U_i(\zeta_{j1}, k) = -\|(x_i + s^*_{ijk}) - \zeta_{j1}\|^2 + \eta_{ij1}, \quad (1)$$

$$U_i(\zeta_{j0}, k) = -\|(x_i + s^*_{ijk}) - \zeta_{j0}\|^2 + \eta_{ij0}, \quad (2)$$

where η_{ij0} and η_{ij0} represent stochastic error terms with mean zero and finite variances, and s^*_{ijk} is the shift in the ideal point of respondent i induced by the endorsement of political actor k for favoring and opposing policy j. Therefore, s^*_{ijk} represents the degree to which political actor k can influence

[6] Formally, the randomization implies $\{Y_{ij}(k)\}_{k=0}^K \perp\!\!\!\perp T_{ij}$ for all i and j and hence we have $\mathbb{E}(Y_{ij}(k) - Y_{ij}(0)) = \mathbb{E}(Y_{ij} \mid T_{ij} = k) - \mathbb{E}(Y_{ij} \mid T_{ij} = 0)$ for any k.

[7] See **?** for an alternative utility model that results in the same statistical model.

respondent i's preference over policy j. Under the control condition of no political endorsement, there is no shift in the ideal point and thus, $s^*_{ij0} = 0$ for all i and j.

Under this framework, the probability that respondent i favors policy j with endorsement by political actor k is given by,

$$\begin{aligned} \Pr(Y_{ij} = 1 \mid T_{ij} = k) &= \Pr(Y_{ij}(k) = 1) \\ &= \Pr(U_i(\zeta_{j1}, k) > U_i(\zeta_{j0}, k)) \\ &= \Pr(\alpha_j + \beta_j(x_i + s^*_{ijk}) > \epsilon_{ij}) \end{aligned} \tag{3}$$

for $k = 0, 1, \ldots, K$ where $\alpha_j = \zeta^2_{j0} - \zeta^2_{j1}$, $\beta_j = 2(\zeta_{j1} - \zeta_{j0})$, and $\epsilon_{ij} = \eta_{ij1} - \eta_{ij0}$. The first equality follows from the randomization of the treatment and implies that we can identify the marginal distribution of the potential responses. We assume that ϵ_{ij} is identically and independently distributed according to some distribution such as the standard normal distribution and the logistic distribution.

Within this framework, we define respondent i's level of support for political actor k given policy j such that a positive value of this parameter implies the respondent favors the political actor and thereby a greater probability that the respondent prefers the policy. Formally, this support parameter is defined as,

$$s_{ijk} = \begin{cases} s^*_{ijk} & \text{if } \beta_j \geq 0 \\ -s^*_{ijk} & \text{otherwise.} \end{cases}$$

Here, the quadratic utility function is convenient because regardless of respondent i's ideal point x_i and the error distribution, a greater value of this support parameter s_{ijk} implies a greater probability that the respondent favors policy j when endorsed by group k.[8] Intuitively, if a viewpoint favoring policy j is located ideologically left of a viewpoint opposing the policy (i.e., $\zeta_{j1} < \zeta_{j0}$ and thus $\beta_j < 0$), the ideal point of a respondent who supports a political actor should be shifted to the left (i.e., $s^*_{ijk} < 0$ and thus $s_{ijk} = -s^*_{ijk}$) if the policy is endorsed by this actor.

In many situations, researchers are interested in modeling how respondents' levels of support for each political actor change as a function of their characteristics. Under our model, this can be accomplished, for example, by modeling support parameters s_{ijk} in the following hierarchical manner,

$$\begin{aligned} s_{ijk} &\overset{\text{indep.}}{\sim} \mathcal{N}(Z_i^\top \lambda_{jk},\ \omega^2_{jk}) \\ \lambda_{jk} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_k, \Phi_k) \end{aligned}$$

[8] Formally, the sign of the first derivative of equation (3) with respect to s^*_{ijk} is the same as the sign of β_j.

where Z_i is an M dimensional vector of respondent i's covariates, λ_{jk}, and θ_k are M dimensional vectors of unknown coefficients, ω_{jk}^2 is a unknown scalar variance, and Φ_k is a unknown $M \times M$ covariance matrix. Since the support parameter is zero for the control condition, i.e., $s_{ij0} = 0$ for all i and j, we also have the restrictions $\lambda_{j0} = 0$ and $\omega_{j0} = 0$ for all j. Similarly, we can model ideal points x_i using the same covariates as follows,

$$x_i \overset{\text{indep.}}{\sim} \mathcal{N}(Z_i^\top \delta,\ \sigma_x^2)$$

where δ is an M dimensional vector of unknown coefficients and σ_x^2 is a known prior variance.

Model Interpretation. What is the key assumption that allows researchers to interpret change in support for a policy as support for an actor who endorses it? As is clear from equations 1 and 2, our model formulation assumes that the endorsement of a policy by an actor may affect the location of respondents' ideal points, but not the locations of favoring and opposing the policy. This assumption will be violated, for example, if respondents are uninformed about the policy and the endorsement provides them with additional information about the content of the policy. Under this alternative scenario, change in support for the policy can be thought of as the result of learning rather than that of expressing support for the actor.

This ambiguity of interpretation is a fundamental problem of endorsement experiments and does not rely upon whether or not researchers employ the proposed statistical model or not. Unfortunately, the observed data cannot tell which interpretation is valid. One possible strategy is to ask each respondent to locate policies on the relevant policy dimension. If we find that the endorsement by an actor does not influence their location, then we may more credibly conclude that change in support for these policies can be attributed to support for the actor. Other possible strategies include choosing policies that are well known to respondents and focusing one's analysis on respondents who are well informed about these policies.

Prior Specification. Finally, our Bayesian model is completed by placing the following independent diffuse prior distributions on all unknown parameters,

$$\begin{aligned}
\alpha_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \\
\beta_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\beta, \sigma_\beta^2) \\
\delta &\sim \mathcal{N}(\mu_\delta, \Sigma_\delta) \\
\theta_k &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\theta, \Sigma_\theta) \\
\omega_{jk}^2 &\overset{\text{i.i.d.}}{\sim} U[0, 10] \\
\text{diag}(\Phi_k) &\overset{\text{i.i.d.}}{\sim} U[0, 10]
\end{aligned}$$

where Φ_k is assumed to be a diagonal matrix.

Generalization to the Ordinal Response Model. To measure the level of support for particular policies, many surveys use ordinal responses (e.g., `strongly oppose`, `oppose`, `neutral`, `favor`, and `strongly favor`) rather than binary responses. Here, we generalize the above measurement model to the ordinal response setup. We use this generalized model in our empirical application.

Suppose that the response variable Y_{ij} is the ordered factor variable taking one of L levels, i.e., $Y_{ij} \in \{0, 1, \ldots, L-1\}$ where $L > 2$. We assume that a greater value of Y_{ij} indicates a greater level of support for policy j. Let ζ_{jl} denote the location of choosing the lth level of support in the single dimensional policy space. Finally, we assume that the elements of $\{\zeta_{jl}\}_{l=0}^{L-1}$ are equally spaced in the policy space in either a descending or ascending order, i.e., $\zeta_{j,l+1} - \zeta_{jl} = \psi_j$ does not vary across levels l although it may vary across policies j. This assumption can be relaxed but we do not pursue such a strategy here to keep the model parsimonious (see below for more discussion on this point).

Then, the utility function can be written as,

$$U_i(\zeta_{jl}, k) \;=\; -\|(x_i + s^*_{ijk}) - \zeta_{jl}\|^2 + \nu_{ijl},$$

which in turn implies the following probability model,

$$\begin{aligned}
\Pr(Y_{ij} \leq l \mid T_{ij} = k) &= \Pr(Y_{ij}(k) \leq l) \\
&= \Pr(U_i(\zeta_{j,l+1}, k) < U_i(\zeta_{jl}, k)) \\
&= \Pr(\alpha_{jl} - \beta_j(x_i + s^*_{ijk}) > \epsilon_{ijl}), \qquad (4)
\end{aligned}$$

for $l = 0, 1, \ldots, L-2$ where $\alpha_{jl} = \zeta^2_{j,l+1} - \zeta^2_{jl} = \psi_j(\zeta_{jl} + \zeta_{j,l+1})$, $\beta_j = 2(\zeta_{j,l+1} - \zeta_{jl}) = 2\psi_j$, and $\epsilon_{ijl} = \nu_{ijl} - \nu_{ij,l+1}$ is assumed to be independently and identically distributed. Note that the relationship between α_{jl} and β_j implies $\alpha_{jl} \leq \alpha_{j,l+1}$ for any j and l. If the assumption of equal spacing between ζ_{jl} and $\zeta_{j,l+1}$ is relaxed, then β_j will need to vary across levels l as well as across policies j. This generalized model corresponds to the random effects factor analytic model with the ordinal response. In the psychometrics literature, this model without random effects is sometimes called a generalized partial credit model (**?**).

3.3 Quantities of Interest

Under the above model, our two basic quantities of interest are defined as,

$$\tau_{jk}(Z_i) \;=\; Z_i^\top \lambda_{jk} \qquad (5)$$

$$\kappa_k(Z_i) \;=\; Z_i^\top \theta_k. \qquad (6)$$

While $\tau_{jk}(Z_i)$ represents the average support level for political actor k with respect to policy j among individuals with characteristics Z_i, $\kappa_k(Z_i)$ denotes the average support level for the political actor

among the same set of individuals without reference to any specific policy. While it is certainly of interest to examine how these two measures of support vary as a function of individual characteristics Z_i, the overall average support levels for each political actor in the population are also important, i.e., $\bar{\tau}_{jk} = \mathbb{E}(\tau_{jk}(Z_i))$ and $\bar{\kappa}_k = \mathbb{E}(\kappa_k(Z_i))$ where the expectations are taken over the population distribution of Z_i.

Although the above measures facilitate the comparison of support levels across different political actors and policies, one important limitation is that their magnitude is difficult to interpret because they are measured in the latent policy space. To address this issue, we standardize levels of support in terms of the posterior standard deviation of ideal points. That is, support for political actors is measured relative to the variation in ideal points among the population. An endorsement impact of 2.0, for example, can now be interpreted as shifting a respondent's ideal point two standard deviations. In the U.S. context, this would be interpreted as having a moderate Republican provide responses akin to those of a moderate Democrat, or vice versa. Thus, the standardized measures provide an important intuition for interpreting their magnitude by placing them on the same scale as ideal points of survey respondents.

3.4 Identification, Estimation, and Inference

Since we have the standard ideal points model for the control group, the local identification results of **?** are directly applicable. Following the convention in the literature, we typically constrain the distribution of estimated ideal points to have mean zero and unit variance. In addition, for global identification, we constrain the sign of at least one element of β_j for a particular policy j (see also **?**). Once the item parameters are identified from the control group, the identification of the average support parameters λ_{jk} is immediate because of randomization. The simulation studies given in Section 5 also show that all relevant parameters are estimated with little bias even in a relatively small sample.

Following many of the recent applications of the item response theory in political science, we use the Markov chain Monte Carlo (MCMC) algorithm to fit the proposed Bayesian measurement model. One can construct the MCMC algorithm using the standard data augmentation scheme with the latent outcome variable Y_{ij}^* underlying the binary or ordered response variable Y_{ij}. One important advantage of Bayesian simulation is that the estimation uncertainty can be easily calibrated using the draws from the posterior distribution.

For actual model fitting, we use the open-source software `JAGS` (**?**, version 2.2.0). As demonstrated in Section 5, the MCMC algorithm implemented through `JAGS` recovers the poster distribution well in the types of data sets we have analyzed. Our replication code is made available as **?** so

that other researchers can apply the proposed methods to their own endorsement experiments.

3.5 Weighting and Post-stratification

Endorsement experiments require a large sample size relative to other approaches especially when the number of political actors and/or the number of policy questions is large. To recoup this loss of efficiency, the use of effective stratification or other sampling techniques at the design stage may be important. Incorporating the resulting survey weights into estimates of support measures obtained from our model is relatively straightforward. Since s_{ijk} represents the level of support at the individual level, survey weights can be directly applied to its estimate when computing the estimated average level of support.

Post-stratification at the analysis stage provides another way of estimating levels of support in the population (**??**). When the population joint distribution of covariates Z_i is available, this can be easily done in our model by properly weighting the estimated average levels of support, $\tau_{jk}(Z_i)$ or $\kappa_k(Z_i)$, according to the distribution of Z_i. We illustrate the use of post-stratification in our analysis of the Pakistan data in Section 4.2.

4 Empirical Results

In this section, we apply the proposed methodology to the data from the Pakistan survey experiment discussed in Section 2. We consider two multilevel/hierarchical models that extend the base model described in Section 3. The first model estimates the levels of support for each militant group at the division level, where the 26 divisions of Pakistan are a now-defunct political unit that is smaller than the four provinces but sufficiently large that we have a reasonable sample in each division. The second model adds individual-level covariates to the first model to examine the variation across individuals. The two models we consider here share the same top level specification, which is given in equation (4) where, based on substantive grounds, we assume that $\beta_j \geq 0$ for all policy questions. This assumption is easily justified for the curriculum reform question and polio vaccination question. Pakistan experts we talked to, our focus group participants, and our enumerators all expected those who support Islamist militancy to oppose them.[9] Both models are also estimated on the set of completed surveys, $n = 5,212$. For model fitting, we run three parallel chains with over-dispersed starting values and monitor the convergence using the standard diagnostic statistic (**?**). After the convergence has been achieved according to this criteria, we keep every thirtieth of the last 60,000

[9]While the majority of our interlocutors expected Islamists to oppose FCR reform and jirgas over the Durand line, opinion was more diverse and some argued Islamists would favor those policies. We have therefore also run the analysis assuming $\beta_j \geq 0$ only for the curriculum reform and polio questions. The results do not change substantively for alternative assumptions.

posterior draws for our inference.

4.1 Estimated Support at the Division Level

Our first model estimates the levels of support for each militant group while assuming that the degree of support is identical across policies. This leads to the following specification,

$$
\begin{aligned}
x_i &\overset{\text{indep.}}{\sim} \mathcal{N}(\delta_{\text{division}[i]}, 1) \\
s_{ijk} &\overset{\text{indep.}}{\sim} \mathcal{N}(\lambda_{k,\text{division}[i]}, \omega_k^2) \\
\delta_{\text{division}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\mu_{\text{province}[i]}, \sigma^2_{\text{province}[i]}) \\
\lambda_{k,\text{division}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\theta_{k,\text{province}[i]}, \Phi_{k,\text{province}[i]})
\end{aligned}
$$

where partial pooling occurs across divisions within each province as well as across individuals within each division.[10]

Figure 2 presents the estimated levels of support for each militant group at the division level that are based on our first model. As discussed in Section 3.3, we standardize our measures of average support for each group by dividing it with the posterior standard deviation of ideal points distribution so that they are easier to interpret than the raw estimates. The plot suggests that the levels of support vary substantially across divisions within each province as well as across provinces. For example, while there is clearly positive support for three of four militant groups in Gujranwala (Punjab), levels of support in Hazara (NWFP) and Nasirabad (Balochistan) are estimated to be negative for all four groups. Moreover, whereas the estimated support tends to be negative within Balochistan and the NWFP, the citizens of Punjab appear to be more favorably disposed towards militant groups.

In contrast to this regional variation, however, the difference in support levels across militant groups within each division is relatively small. Although some variations exist in such divisions as Sibi and Zhob in the Balochistan province where *firqavarana tanzeems* attracts relatively large negative support, the group's 90% confidence interval overlaps substantially with those of the other militant groups.

4.2 Estimated Effects of Individual Level Covariates and Use of Post-stratification

The second model we consider is the same as the first model presented in Section 4.1 except that we incorporate individual level covariates in order to answer substantive questions and gain statistical efficiency. Here, we use gender, education, income, and whether respondents are from an urban area.

[10] We also fitted another model that allows the intercepts α_{jl} in equation (4) to vary across divisions and place an independent and diffuse prior. The results are similar to the ones presented here.

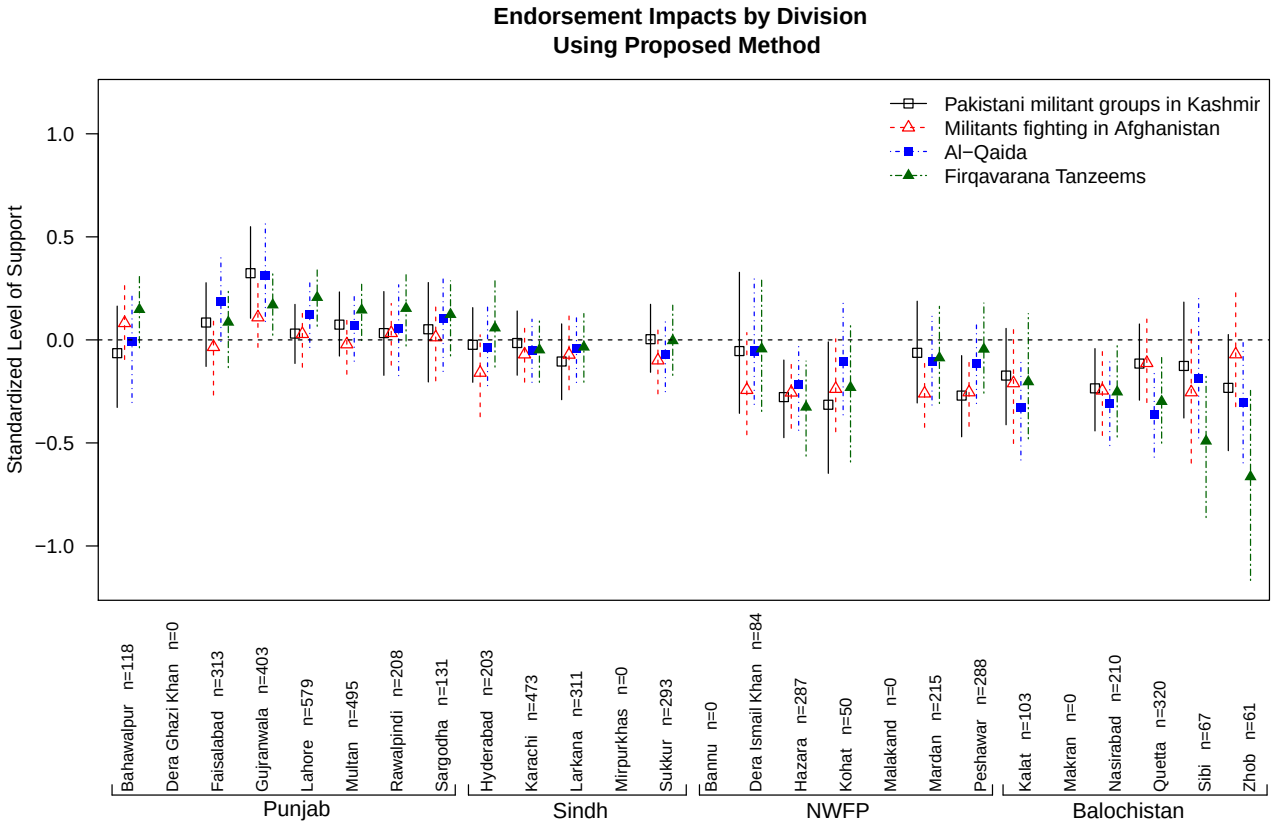

Figure 2: Estimated Division Level Support for Each Militant Group. The vertical axis represents the estimated levels of support, which are standardized in terms of the posterior standard deviation of ideal points distribution. The symbols represent the point estimates and the vertical lines are 90% Bayesian confidence intervals. In this model, the degree of support is assumed to be constant across policies. The plot suggests that the levels of support varies substantially across divisions and provinces. In contrast, little variation is found across different militant groups within each division.

[11] Since the joint population distribution of these variables is available from the recent Pakistan Social and Living Standards Measurement Survey (PSLM), we can post-stratify the resulting estimates on these covariates to compute the division level estimates of support level that adjust for remaining discrepancies between the sample and the population. This model with individual covariates are

[11] Following the work of **?**, we code respondents' income as 1 if their household income was within the bottom quartile for their province, as 2 if their household income was within the middle quartiles for their province, and as 3 if their household income was within the top quartile for their province. We code respondents' education as 1 if illiterate, as 2 if they have only primary school education, as 3 if they have middle school education, as 4 if they have matric education, and as 5 if they have intermediate (f.a/f.sc) or higher education.

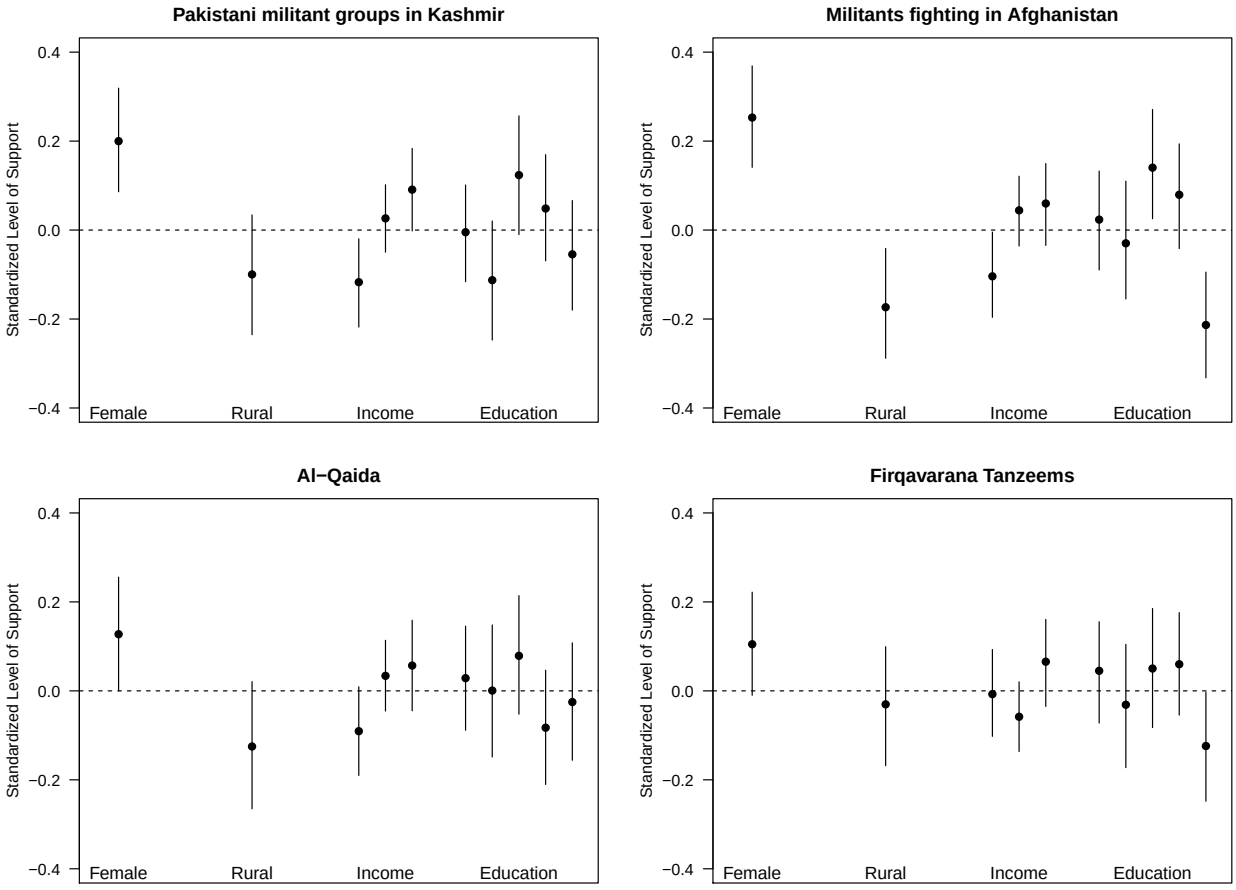

Figure 3: Estimated Effects of Individual Covariates on Levels of Support. The vertical axis represents the estimated effects of individual level covariates on levels of support, which are standardized in terms of the posterior standard deviation of ideal points distribution. The symbols represent the point estimates and the vertical lines are 90% Bayesian confidence intervals. The plot shows that after accounting for regional differences in support levels, women are slightly more supportive of each group and respondents from rural areas as marginally less supportive of each group. There appears to be no clear patter of support across income or education groups after accounting for geographic variation.

given by,[12]

$$
\begin{aligned}
x_i &\overset{\text{indep.}}{\sim} \mathcal{N}(\delta_{\text{division}[i]} + Z_i^\top \delta^Z,\ 1) \\
s_{ijk} &\overset{\text{indep.}}{\sim} \mathcal{N}(\lambda_{k,\text{division}[i]} + Z_i^\top \lambda_k^Z,\ \omega_k^2) \\
\delta_{\text{division}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\mu_{\text{province}[i]}, \sigma^2_{\text{province}[i]}) \\
\lambda_{k,\text{division}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\theta_{k,\text{province}[i]}, \Phi_{k,\text{province}[i]}).
\end{aligned}
$$

Figure 3 presents the results based on this model. The figure shows that individual covariates explain only a small portion of the variance in support levels once we take into account regional

[12]Again, we also fitted a model that allows the intercepts α_{jl} to vary across divisions. The results are similar to the ones we obtain here.

differences in support across divisions. It appears that women are slightly more supportive of each group and respondents from rural areas are marginally less supportive of each group. However, the predictive power of these covariates is small relative to the recovered estimates from geographic divisions. As such, the division-level estimates produced by employing post-stratification (not shown) are little-changed from the estimates shown in Figure 2.

4.3 Substantive Discussion of Empirical Results

One way to understand the result of our Bayesian measurement model is to view the division-level support estimates overlaid on a map of Pakistan, as shown in Figure 4. Viewing the data in this manner highlights the spatial patterns of support for militant organization that are not necessarily clear from tables and figures. Examining the spatial variation is especially important given our finding that individual covariates matter little for predicting the levels of support once regional differences are taken into account.

The most obvious pattern we see is that support for groups clusters coherently within provinces, with support for al-Qa'ida and the militants fighting in Kashmir clustered in Gujranwala in north-eastern Punjab. The second most obvious pattern is that support for the Afghan Taliban is highest in the parts of the country furthest from Afghanistan. Put slightly differently, it appears that, on average, Pakistanis most removed from the consequences of the war in Afghanistan are most supportive of the non-state actors fighting that war. Critically, Pakistanis in the areas where senior Taliban leadership are thought to reside (Quetta and Zhob) are least tolerant of their organization. Finally, we see that support for the *firqavarana tanzeem* is spread evenly throughout Punjab, not concentrated in Bahawalpur as some have argued.

More specifically, our approach reveals several substantively important patterns. First and foremost, there is clear variation in the intensity of support across divisions within provinces. In the Punjab where respondents are generally weakly positive towards most groups, the intra-provincial variation does not match what one would expect from following ongoing policy discussions. Strikingly, we only find a significant positive endorsement effects for one group, the *firqavarana tanzeems*, in Bahawalpur, which has been widely cited as the locus of militancy within the Punjab.[13] If anything Bahawalpur has the lowest average levels of support in this province. Instead, we find strong positive endorsement effects for three of four groups in Gujranwala which has received much less attention from policy makers.[14] We also see positive endorsement effects for the firqavarana tanzeem in four of seven divisions within Punjab which is consistent with the idea that many of the sectarian

[13]Bahawalpur district, the center of the division, was listed as one of the 26 USAID focus districts under a now-defunct plan to direct aid to the areas of Pakistan posing the greatest threat of militancy.

[14]For example, none of the districts in Gujranwala division were among the 26 USAID focus districts.

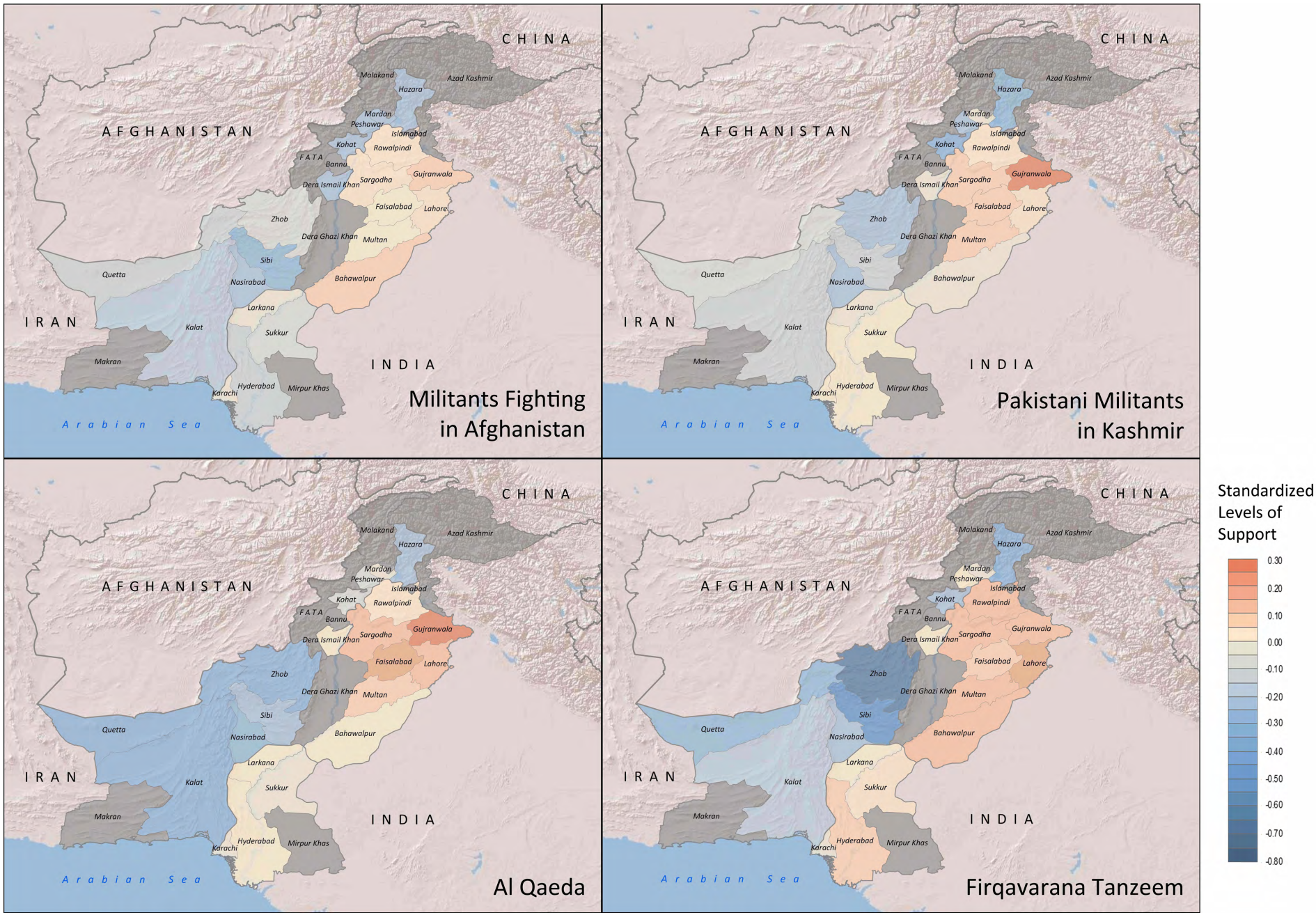

Figure 4: Map of Estimated Standardized Levels of Support for Four Militant Groups in Pakistan. The red represents the estimated positive support while the blue represents the estimated negative support. The dark grey represents the areas where the survey was not conducted and therefore no estimates are available. The estimated levels of support are standardized using the standard deviation of the (posterior) ideal points distribution as a unit.

groups draw their support from the Punjab.

Second, areas of Pakistan that suffered a great deal of militant violence before April 2009 show the most negative endorsement effects. Hazara, Kohat, Nasirabad, and Peshawar all suffered particularly heavily from militant violence in early 2009. The results for Quetta are particularly interesting in that this region has long been a safe-haven for Afghan Taliban leaders but also suffered heavily from sectarian violence in early 2009. As one might therefore expect, groups espousing explicitly sectarian political goals (al-Qa'ida and the *firqavarana tanzeem*) fare poorly in public opinion in Quetta. The policy point here is a subtle one in so far as levels of militant activity should not be taken to indicate popular support. Indeed, our analysis suggests that quite the opposite may be true.

To probe this insight further, we combine data on attitudes from the survey with panel data on militant violence. The panel data were constructed by geo-referencing the 5,601 terrorist incidents recorded in Pakistan since 2004 by the National Counterterrorism Center's Worldwide Incident Tracking System (WITS).[15] The WITS data record the location and consequences of all attacks that involve "...premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agent," but do not list the party responsible for the vast majority of incidents. We examine the bivariate relationship between the estimated support levels and the total number of terrorist incidents from March 2008 through March 2009 (12 months prior to the implementation of the survey).

The left column of Figure 5 shows that the impression one takes away from looking at the spatial distribution of violence holds in the data. There is a clear negative relationship between total violence in the year leading up to our survey and the average level of support for groups. This relationship is strongest for militants focused on Kashmir and the Afghan Taliban, and the correlations between total incidents and support are strongly negative for all four groups.[16] Strikingly, no division that experienced more than 50 incidents in the year before the survey was fielded shows positive support for any militant group.

How do the results based on our proposed model differ from the standard statistical methods that are often used to analyze endorsement experiments? While the direct comparison is somewhat difficult because two models are completely different, we investigate the substantive differences in the resulting estimates of support levels for militant groups. As a standard method, we use the ordered probit model where the responses are regressed on the policy indicator variables, the division indicator variables, the treatment variables and the interaction of the division indicator variables

[15] We thank Christine Fair and Hamzah Saif for providing this data set.

[16] The correlations are: -0.57 (p-value $= .006$) for the Pakistani militant groups in Kashmir, -0.59 (p-value $= .005$) for the Afghan Taliban, -0.47 (p-value $= .032$) for Al-Qa'ida, and -0.41 (p-value $= .062$) for the *firqavarana tanzeem*.

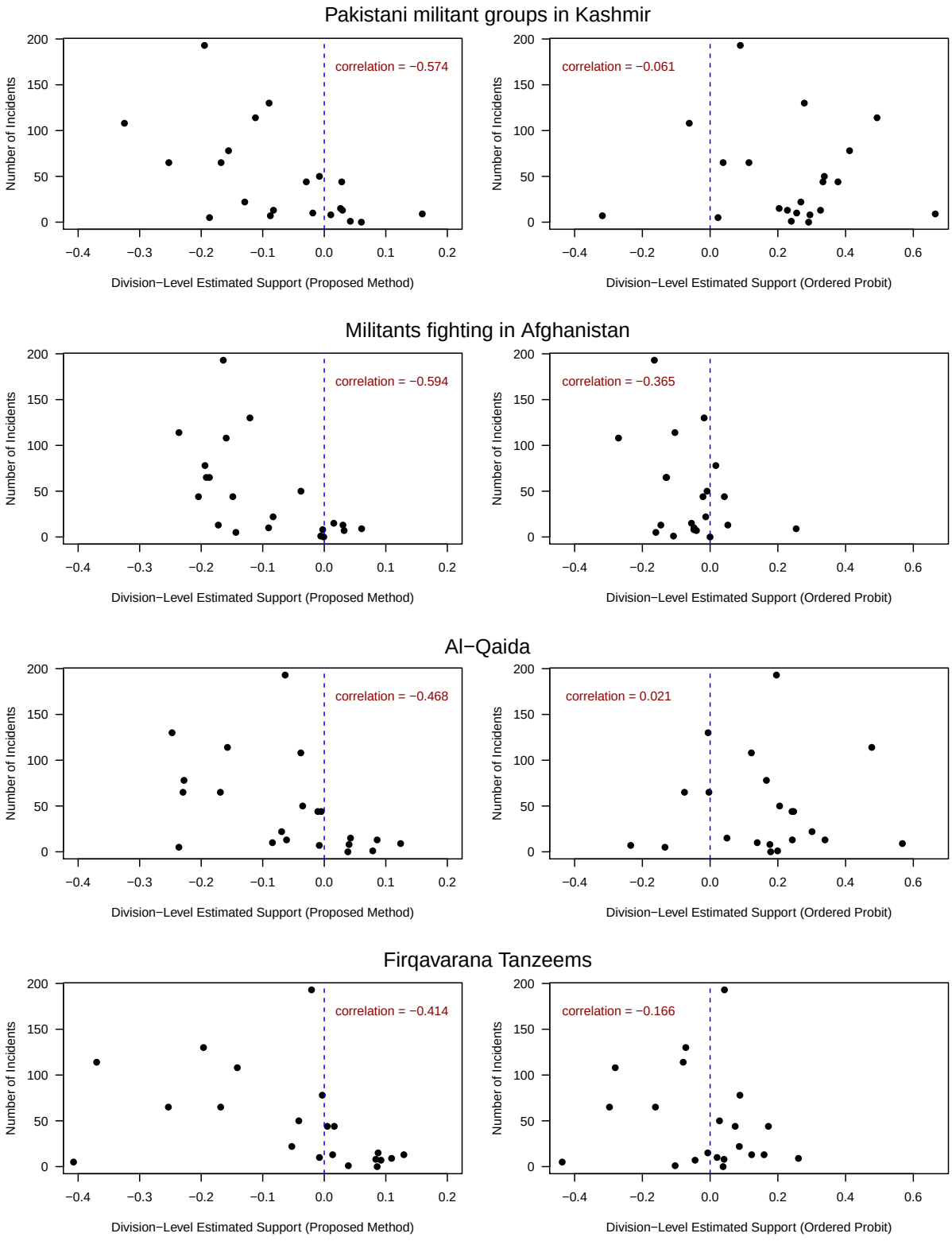

Figure 5: Bivariate Relationship Between Levels of Support for Four Militant Groups in Pakistan and Incidents of Violence within Divisions. In the first column, the horizontal axis represents the estimated level of standardized support within the division for each militant group obtained from the proposed method. In the second column, the horizontal axis represents the analogous estimate obtained from using a standard ordered probit model. The vertical axis represents the total number of terrorist incidents in each division from March 2008 through March 2009 (12 months prior to the implementation of the survey) according to the Worldwide Incident Tracking System (WITS) data. For each group, there exists a negative correlation between the total violence in the area and the estimated average level of support from our model. This correlation is much weaker when average level of support is estimated using the standard ordered probit model.

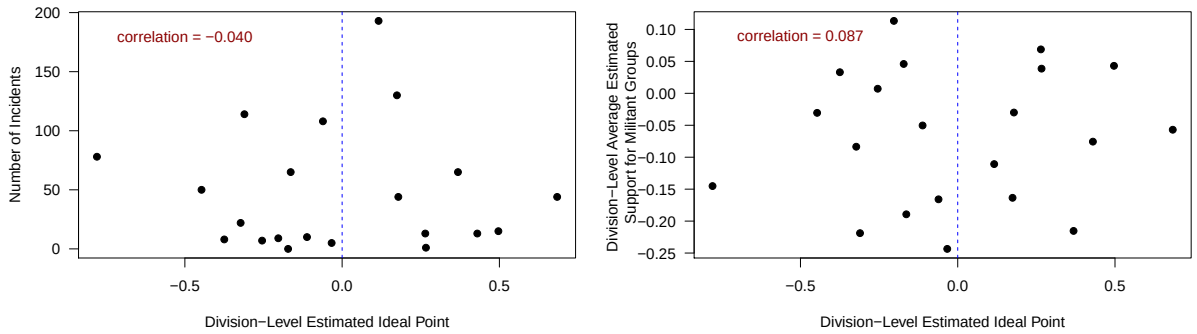

Figure 6: Bivariate Relationship Between Ideology, Incidents of Violence and Average Levels of Support for Militant Groups in Pakistan within Divisions. In the both plots, the horizontal axis represents the estimated ideology standardized within the division. In the first column, the vertical axis represents the total number of terrorist incidents in each division from March 2008 through March 2009 (12 months prior to the implementation of the survey) according to the Worldwide Incident Tracking System (WITS) data. In the second plot, the vertical axis represents the estimated level of standardized support within the division, averaged across the four militant groups. From both plots, we see no relationship between division ideology and either division violence or division support for the militant groups.

with the treatment variables, with standard errors clustered by respondent. This simple model assumes that the coefficients for the treatment variables are identical across policy questions and that threshold parameters of response categories are equally spaced across policy questions. We then interpret the coefficients for the treatment variables as scalar measures of support level for militant groups.

The right column of Figure 5 shows the relationship between the same terrorist incidence data and the estimated coefficients from the ordered probit model. In contrast to the estimates based on our model, there is little relationship between total violence in an area and estimated average support for each militant group based on a standard ordered probit. In each case, the correlations are considerably lower, and none reach standard levels of statistical significance (p-values of $.793, .104, .927, \text{and} .473$). Thus, analysis using a standard ordered probit model would miss a substantively critical result discovered by our model.

Finally, one concern with these results is that ideological differences across divisions may explain the correlation between terrorist incidents and support levels for militant groups. For example, if ideologically liberal areas experience more terrorist attacks and residents of those regions give less support for militant groups, then the negative correlation between violence and support we found above is spurious. To address this possibility, we examine how division-level ideal points in policy space relate to both total violence in an area, and to average support for the militant groups in that area. As seen in Figure 6, there exists no clear relationship between division-level ideal points and total violence. Similarly, there is no strong correlation between ideal points and support at the division level (the corresponding individual level correlation is also weak, ranging between -0.08 and

-0.10). While this in itself is an important substantive finding, the key value of estimating ideal points for this particular analysis is that they permit the diagnostics, which provide confidence that our core empirical results are not an artifact of the correlation between violence and ideal points over the underlying policies.

Overall, these results provide evidence for the argument that support for Islamist militancy in Pakistan is concentrated in divisions where there is little violence and in the most economically developed and politically powerful part of the country such as northern Punjab.

5 Simulation Studies

We assess the statistical properties of our model as implemented through `JAGS` under two circumstances. The first set of simulations is based on the data generating processes that closely follow the observed data of our Pakistan survey experiment. This allows us to examine the reliability of our estimates given in Section 4. The next set of simulations is based upon sample sizes and data generating processes that are more typical among survey experiments in political science. This provides a way to examine whether our proposed method is more broadly applicable. In both sets of simulations, we investigate the frequentist statistical properties of the Bayesian hierarchical measurement models used in our analysis of the Pakistani data, i.e., the "Division Model" and the "Division Model with Individual Covariates" as specified in Sections 4.1 and 4.2, respectively.

Below, we report the bias of point estimates (based on posterior mean), the coverage of 90% Bayesian confidence intervals, and the statistical power of $\alpha = 0.1$ level Bayesian hypothesis test. These statistics are computed over 100 simulations. For each simulation run, the model is fitted with three parallel chains and random starting values. The number of iterations is roughly 60,000. Convergence was diagnosed by two criteria; the values of the $\widehat{R}$ statistics and visual inspection of trace and autocorrelation function plots. For expository purpose, we do not provide the detailed information about our simulations (e.g., the true values of the model parameters). Such information and all other details about the data generating process is made available in the replication archive (**?**).

5.1 Monte Carlo Evidence based on the Pakistan Data

The first set of simulations is based upon the data generating processes that closely follow the observed data from our Pakistan survey experiment. The sample size is $5,212$, which is larger than typical political science survey experiments. The non-issue-specific parameters (i.e., $\delta_{division}$, $\lambda_{k,division}$, ω_k, δ^Z, λ_k^Z) are kept constant across all simulations. Throughout the simulations, we use the same distribution of covariates and sample sizes as the Pakistan data.

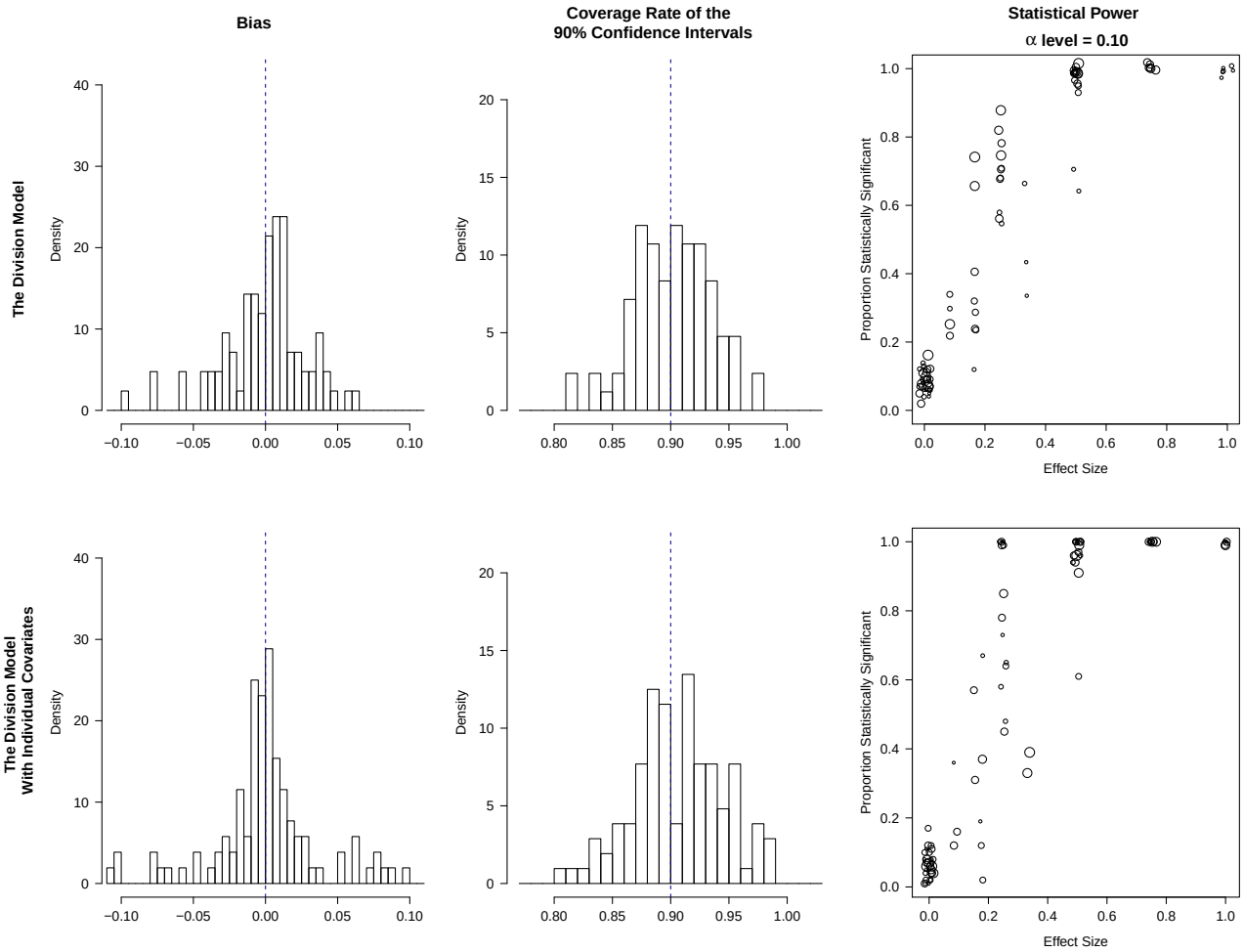

Figure 7: Simulation Results based on the Pakistan Data. Across both models, the distribution of bias across parameters (left column) is centered around zero is small in magnitude relative to the standard deviation of respondents' ideal points. The coverage rates of the 90% Bayesian confidence intervals across parameters (middle column) are also centered around the nominal rate. Finally, the statistical power of the proposed models (represented as the proportion of statistically significant estimates in the right column) is quite high and increase rapidly with effect size. Moreover, we observe that the statistical power is a function of sample size within each division, which is reflected by the size of circles.

For each simulation, we sample respondents' ideal points and potential endorsement impacts for all groups and policies according to our model (conditional on their covariates when appropriate). After splitting the sample into equal halves control group and treatment group, the sequence of endorsers is randomized for the treatment group. We then generate the responses to a set of questions, each of which has a random set of cutpoints and discrimination parameter according to our model specification. In addition, the true values of the model parameters are chosen such that these responses were to fall within a plausible range. This constitutes the generic data generating process.

As with any set of simulations testing a statistical model, we would ideally observe three criteria. First, there should be no bias in the recovered parameter estimates. Second, the recovered confidence

intervals should be of appropriate width. Confidence intervals that are too short will lead us to have inappropriately high confidence in our point estimates, while intervals that are too long will leave us unable to make clear distinctions from the data. Finally, given the traditional focus on hypothesis testing, we want our tests to have large statistical power if possible. That is, we would like to detect the departure from the null hypothesis whenever it exists.

The results from this set of simulation studies are summarized in Figure 7. For the Division Model we observed an average bias of -0.00735 and an average root mean squared error (RMSE) of 0.138. The coverage rate of the 90% confidence intervals was 0.887 with a standard deviation of 0.038. Finally, the statistical power of the proposed models (represented as the proportion of statistically significant at the $\alpha = 0.1$ level in the right column) is quite high and increase rapidly with effect size and division level sample size (represented by the size of circles in the top right corner graph). Moreover, the Type I (false positive) error rate was 0.0992, close to the nominal level of $\alpha = 0.1$.

The results for the Division Model with Individual Covariates are equally impressive in terms of all three criteria, except that the statistical power may be low for some parameters. Overall, these models seem to recover the true parameter values efficiently and their confidence intervals have appropriate coverage rates.

5.2 Monte Carlo Evidence with Varying Sample Sizes

Most political science datasets are much smaller than the Pakistan dataset, and more often the responses to survey questions have central response distributions. To give a better understanding of the performance of our model in typical situations, we conduct simulation studies with varying sample sizes. In particular, we repeat the data generating process as above, but with three exceptions. First, the number of geographic units was reduced from 26 divisions to 16 divisions, with equal sample sizes across divisions. Second, the sample size is reduced considerably to sizes 1000, $1,500$, and $2,000$. Third, the distribution of responses has a stronger central tendency on the $1-5$ scale.

Figure 8 presents the simulation results based on all three models. These models perform reasonably well with small bias (left column) and appropriate coverage probabilities (middle column). In particular, as the sample size increases, the bias gets smaller and the coverage rate improves. As expected, the statistical power (right column) does decrease as the sample size gets smaller. However, even when the sample size is as small as 1,000, the statistical power is reasonable; we reject the null hypothesis of zero effect about half of the time for the parameter whose true value is around 0.4.

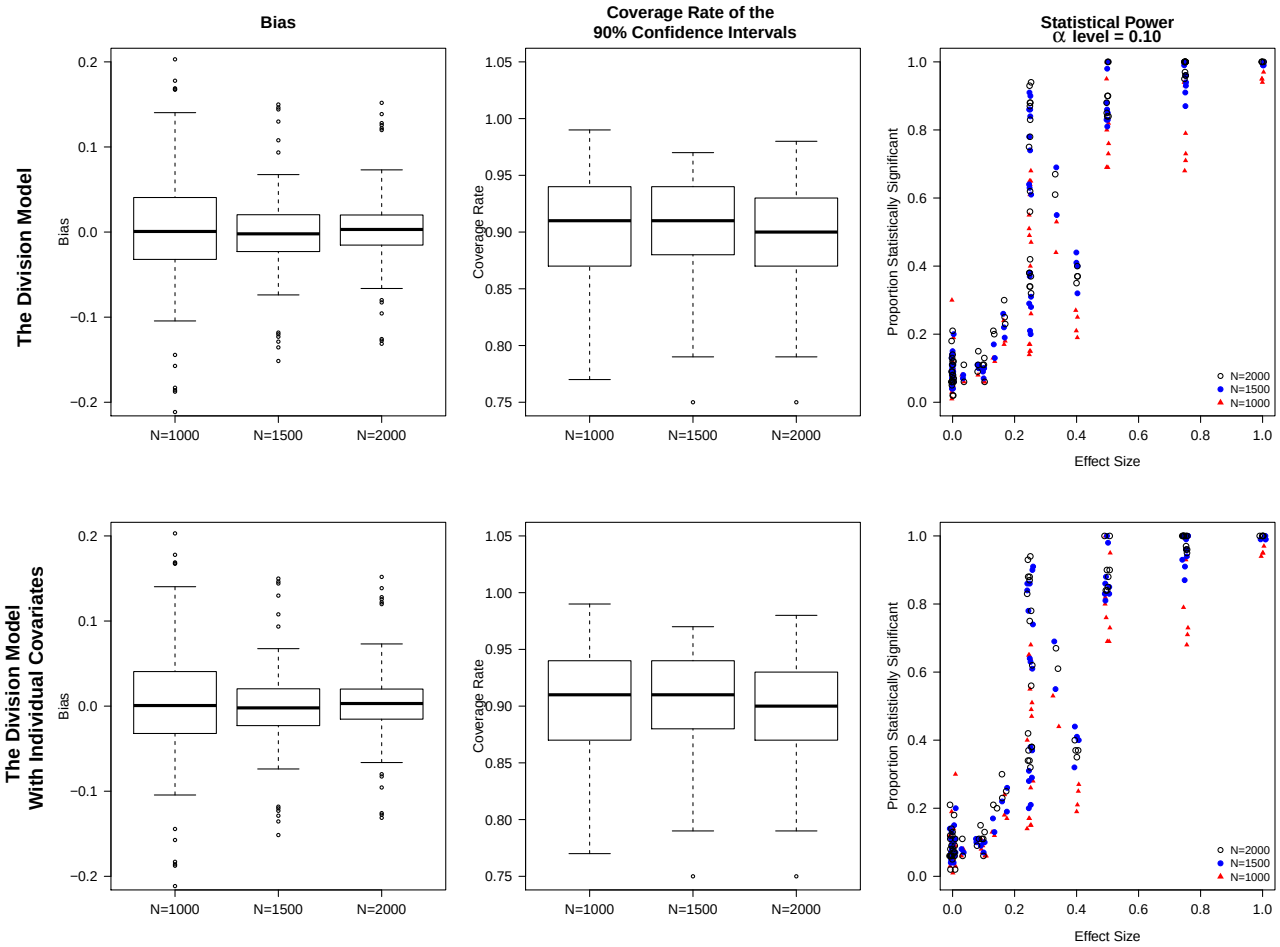

Figure 8: Simulation Results with Varying Sample Sizes. Each column represents a criteria by which the simulation results should be judged, while each row represents a different model. As with the previous set of simulations, we find little evidence of bias or inadequate coverage. As expected, however, the statistical power does decrease as the sample size gets smaller.

5.3 Summary of Simulation Results

Our ability to recover precise, unbiased parameter estimates gives us confidence that our estimation procedure is well-suited for endorsement experiments. While no set of simulations can cover all plausible situations a researcher might face, our model performs well across a broad range of scenarios. Table 1 shows that the proposed models have good frequentist properties on several criteria. For each set of simulations, the average bias (first column) across all parameters of interest is negligible. The average RMSE (second column) is larger than one might hope. But the coverage statistics of the 90% confidence intervals (third column) are of appropriate size, which suggests that the errors are not necessarily problematic. Moreover, the Type I (false positive) error rate is consistently close to the nominal level of $\alpha = 0.1$ (last column). When compared to ordinal logistic regression and other standard estimation techniques, our model performs well. For example, using ROC curves to

Model	N	Average Bias	Average RMSE	Average Coverage	Type I Error
Division	5212	−0.001	0.137	0.907	0.063
Division w/ Individual Covariates	5212	0.000	0.123	0.911	0.066
Division	1000	0.006	0.251	0.888	0.081
Division	1500	0.003	0.191	0.900	0.080
Division	2000	−0.003	0.178	0.890	0.086
Division w/ Individual Covariates	1000	−0.001	0.224	0.900	0.069
Division w/ Individual Covariates	1500	−0.003	0.173	0.901	0.071
Division w/ Individual Covariates	2000	0.004	0.163	0.894	0.071

Table 1: Simulation Summary Statistics. For each set of simulations, the average bias across all parameters of interest is negligible. Upon first glance the average RMSE is not as small as one might hope. But the coverage statistics of the 90% confidence intervals are of appropriate size, which suggests that the errors are not necessarily problematic. Moreover, the Type I (false positive) error rate is consistently close to the nominal level of $\alpha = 0.1$.

asses in-sample and out-of-sample prediction rates, our model is equally or slightly more efficient while remaining unbiased (not shown).

6 Concluding Remarks and Suggestions for Applied Researchers

Endorsement experiments are a survey methodology designed to ascertain levels of political support for socially sensitive actors. While their effectiveness must be empirically validated more extensively in the future, endorsement experiments may provide researchers with one way to indirectly ask sensitive questions, thereby avoiding the problems of social desirability bias and item non-response. The Pakistan experiment analyzed in this paper is such an instance where directly asking about support for militant groups in many parts of the country would have raised safety concerns for survey respondents and enumerators as well as possibility of bias and high non-response.

In this paper, we develop a new Bayesian hierarchical measurement model to analyze endorsement experiments by appropriately combining multiple policy questions and yielding estimates of support within the item response theory framework so that they are interpretable. Through our empirical application, we show that our Bayesian measurement model can effectively recover subtle regional differences in support levels, yielding new substantive results which standard methods fail to find. The simulation studies demonstrate that across different sample sizes the proposed methodology has good frequentist properties; low bias, confidence intervals with good coverage, and relatively large statistical power.

Applying our model to the data from Pakistan revealed a number of key facts that were not apparent from the commonly used statistical models. We showed that attitudes towards militant groups are highly regionally clustered and that support is generally lower where there have been a

large number of militant attacks. We also find that individual level variables that are the focus of much policy attention — many of the aid programs directed at addressing the threat from Islamist militancy focus on increasing education and income, for example — matter little once regional differences in support levels are taken into account.

Despite these advantages, it is important to recognize the limitations of endorsement experiments and the proposed methodology. In particular, as discussed in Section 3.2, inferring level of political support from endorsement experiments requires the critical assumption that difference in respondents' support for policies induced by endorsement can be attributed to support for the group rather than respondents' learning about underlying policies. Since the observed data alone cannot verify this assumption, we provide the following practical recommendations for applied researchers who are planning to use endorsement experiments in their research.

- Ask survey respondents to locate each policy before assessing support and examine whether endorsement affects locations of policies. If it does, the finding suggests that respondents in the treatment group are learning about policies, making it difficult to infer support level from endorsement experiments.

- Choose policies that are well-known among respondents and are on the same uni-dimensional policy space. This minimizes the possibility that endorsement provides new information about these policies and avoids the statistical model with a high-dimensional policy space that is difficult to estimate and interpret.

- Measure the level of political knowledge each respondent has about the selected policies. Such a measure allows researchers to focus their analysis on the subset of respondents who are less likely to learn about policies themselves from endorsement.

Using surveys to measure political support for socially sensitive actors creates many challenges with respect to the design, implementation, analysis, and interpretation of surveys. While there is no perfect solution, endorsement experiments provide a possible way to reduce bias resulting from social desirability and nonresponse. We show that it is possible to analyze endorsement experiments within the item response theory framework and obtain measures of support levels on the same scale as ideal points of survey respondents. The proposed model efficiently combines responses across different policy questions and shows how much the endorsement by each actor can shift the ideal points of respondents on the common ideological dimension. Equipped with the new tools, we hope that researchers can now effectively apply endorsement experiments to measure political support when direct questioning is either impossible or unlikely to yield reliable measures.