Eliciting Truthful Responses to Sensitive Survey Questions

Kosuke Imai

Princeton University

Collaborators:
Graeme Blair, Will Bullock, Jason Lyall (Yale)
Jacob Shapiro, Carlos Velasco Rivera, Yuki Shiraito

March 5, 2012

Motivation

- Survey is used widely in social sciences
- Validity of survey depends on the accuracy of self-reports

- Sensitive questions $\Longrightarrow$ social desirability, privacy concerns
- Racial prejudice, corruption, support for political actors
- Lies and nonresponses $\Longrightarrow$ potential bias

- Survey "experiments" as a solution
  1. Randomization: Randomized response method
  2. Aggregation: List experiment (item count technique)
  3. Cueing: Endorsement experiment

- Goal of the project:
  1. Develop statistical methods for analyzing these experiments
  2. Develop recommendations for designing these experiments
  3. Measure support for militants and foreign troops

List Experiment: An Example

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the control group

```
Now I'm going to read you three things that sometimes
make people angry or upset. After I read all three,
just tell me HOW MANY of them upset you. (I don't
want to know which ones, just how many.)

(1) the federal government increasing the tax on
gasoline;
(2) professional athletes getting million-dollar-plus
salaries;
(3) large corporations polluting the environment.
```

List Experiment: An Example

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the treatment group

Now I'm going to read you four things that sometimes make people angry or upset. After I read all four, just tell me HOW MANY of them upset you. (I don't want to know which ones, just how many.)

(1) the federal government increasing the tax on gasoline;
(2) professional athletes getting million-dollar-plus salaries;
(3) large corporations polluting the environment;
(4) a black family moving next door to you.

Challenges and Progress for List Experiments

- List experiment is becoming popular
- Standard practice: Use difference-in-means to estimate the proportion of those who answer yes to sensitive item
- Getting more out of list experiments:
  1. Who are more likely to answer yes?
  2. Who are answering differently to direct and indirect questioning?
  3. Can we study multiple sensitive items in one survey?
  4. Can we detect failures of list experiments?
  5. Can we correct violations of key assumptions?
- We have developed:
  1. efficient multivariate regression analysis methodology
  2. statistical tests to detect failures of list experiments
  3. statistical methods to model deviations from the key assumption

Identification Assumptions

1. Randomization of the Treatment
2. No Design Effect: The inclusion of the sensitive item does not affect answers to control items
3. No Liars: Answers about the sensitive item are truthful

Under these assumptions, difference-in-means estimator is unbiased

New Multivariate Regression Estimators

- Notation:
  - J: number of control items
  - N: number of respondents
  - T_i: binary treatment indicator (1 = treatment, 0 = control)
  - X_i: pre-treatment covariates
  - Y_i: outcome variable

- The nonlinear least squares regression model:

$$Y_i = \underbrace{f(X_i, \gamma)}_{\textit{control items}} + \underbrace{T_i \cdot g(X_i, \delta)}_{\textit{sensitive item}} + \epsilon_i$$

- Difference-in-means: no covariate
- Linear model: $f(x, \gamma) = x^\top \gamma$ and $g(x, \delta) = x^\top \delta$

- Logit model: $f(x, \gamma) = J \cdot \text{logit}^{-1}(x^\top \gamma)$ and $g(x, \delta) = \text{logit}^{-1}(x^\top \delta)$
- Two-step estimation with appropriate standard error

Extracting More Information from List Experiments

- Define a type of each respondent by
  - total number of yes for control items $Y_i(0)$
  - truthful answer to the sensitive item Z_i^*
- A total of $(2 \times (J + 1))$ types
- Example: three control items ($J = 3$)

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0)	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	(0,1) (1,0)	(1,1) (1,0)
0	(0,0)	(0,1) (0,0)

Extracting More Information from List Experiments

- Define a type of each respondent by
  - total number of yes for control items $Y_i(0)$
  - truthful answer to the sensitive item Z_i^*
- A total of $(2 \times (J + 1))$ types
- Example: three control items $(J = 3)$

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0)	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	~~(0,1)~~ ~~(1,0)~~	(1,1) ~~(1,0)~~
0	~~(0,0)~~	~~(0,1)~~ ~~(0,0)~~

- *Joint distribution* of $(Y_i(0), Z_i^*)$ is identified

The Maximum Likelihood Estimator

- Model for sensitive item as before: e.g., logistic regression

$$\Pr(Z^*_{i,J+1} = 1 \mid X_i = x) \;=\; \text{logit}^{-1}(x^\top \delta)$$

- Model for control items given the response to sensitive item: e.g., binomial or beta-binomial logistic regression

$$\Pr(\, Y_i(0) = y \mid X_i = x, Z^*_{i,J+1} = z) \;=\; J \times \text{logit}^{-1}(x^\top \psi_z)$$

- Difficult to maximize the resulting complex likelihood function
- Develop the EM algorithm for reliable estimation

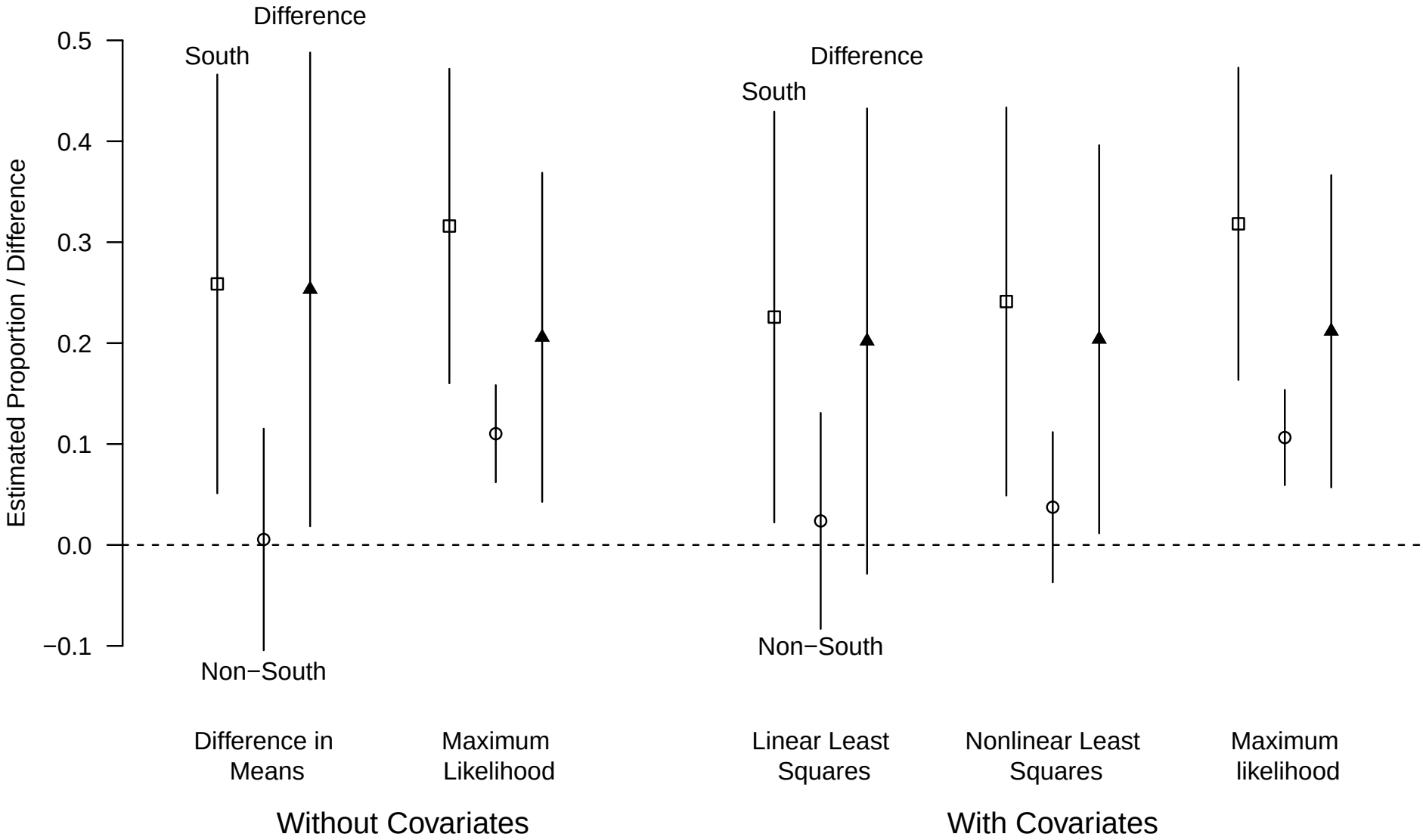
Empirical Application: Racial Prejudice in the US

- Kuklinski *et al*. (1997 JOP): Southern whites are more prejudiced against blacks than non-southern whites – no "New South"
- The limitation of the original analysis:

  So far our discussion has implicitly assumed that the higher level of prejudice among white southerners results from something uniquely "southern," what many would call southern culture. This assumption could be wrong. If white southerners were older, less educated, and the like – characteristics normally associated with greater prejudice – then demographics would explain the regional difference in racial attitudes

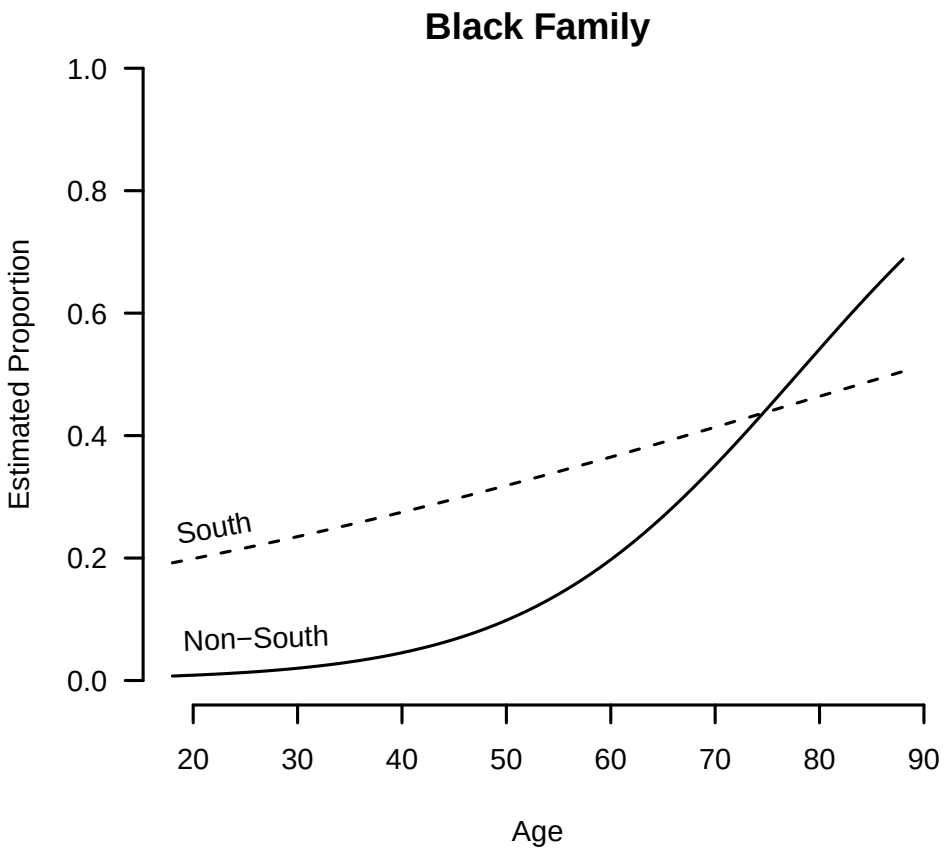
- Need for a multivariate regression analysis

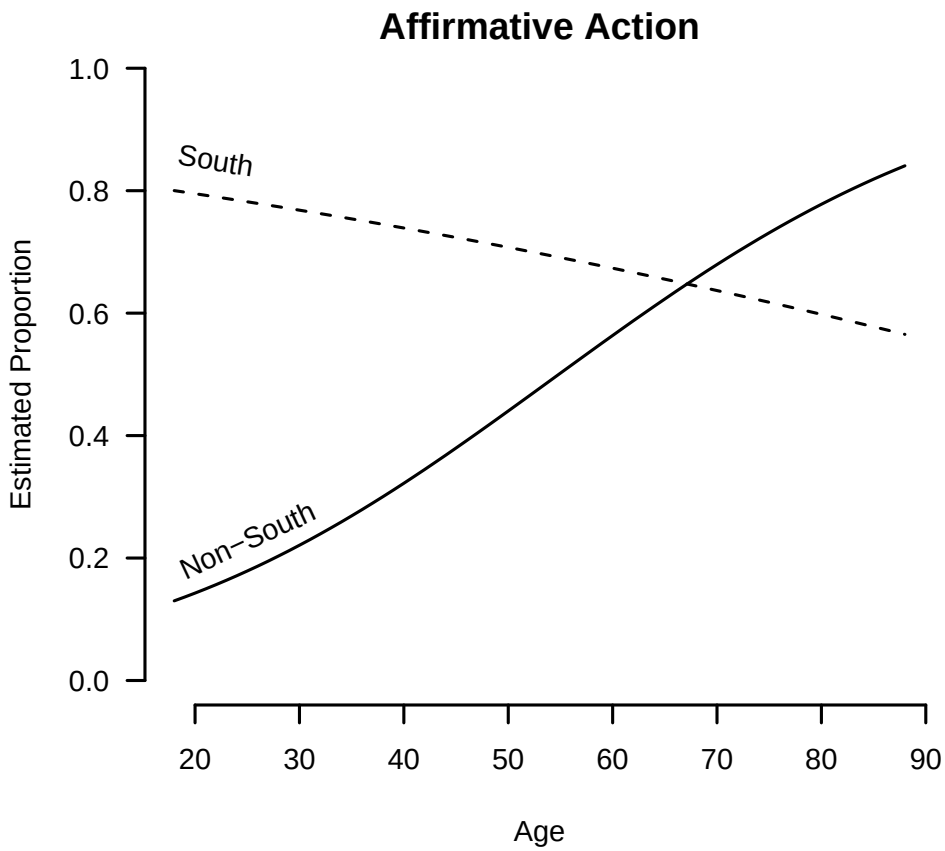
Estimated Proportion of Prejudiced Whites



- MLE yields more efficient estimates

Generational Changes in South and Non-South





- Age is important even after controlling for gender and education
- Gender is not, contradicting with the original analysis

When Can List Experiments Fail?

- No Design Effect
- Respondents may evaluate control items relative to sensitive item

- No Liars
- Ceiling effect: too many yeses for control items
- Floor effect: too many noes for control items

- Question: Can these failures be addressed statistically?

Hypothesis Test for Detecting List Experiment Failures

- Under the null hypothesis of no design effect and no liars, we expect proportions of all "types" to be properly estimated
- Alternative hypothesis: *At least* one is negative
- Correction for multiple testing

	Observed Data				Estimated Proportion of Respondent Types			
	Control		Treatment					
Response	counts	prop.	counts	prop.	$\hat{\pi}_{y0}$	s.e.	$\hat{\pi}_{y1}$	s.e.
0	8	1.4%	19	3.0%	3.0%	0.7	−1.7%	0.8
1	132	22.4	123	19.7	21.4	1.7	1.0	2.4
2	222	37.7	229	36.7	35.7	2.6	2.0	2.8
3	227	38.5	219	35.1	33.1	2.2	5.4	0.9
4			34	5.4				
Total	589		624		93.2		6.8	

- p-value $= 0.022$

Modeling Ceiling and Floor Effects

- Potential liars:

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0) (3,1)*	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	(0,1) (1,0)	(1,1) (1,0)
0	(0,0) (0,1)*	(0,1) (0,0)

- Proposed strategy: model ceiling and/or floor effects under an additional assumption
- Identification assumption: conditional independence between items given covariates
- ML regression estimator can be extended to this situation
- A similar strategy applicable to design effects

Practical Suggestions for List Experiments

- Suggestions for analysis:
  1. Estimate proportions of types and test design effects
  2. Conduct multivariate regression analyses
  3. Investigate the robustness of findings to ceiling and floor effects

- Suggestions for design:
  1. Select control items to avoid skewed response distribution
  2. Avoid control items that are ambiguous and generate weak opinion
  3. Conduct a pilot study and maximize statistical power

- Open-source software:
  - R package `list: Statistical Methods for the Item Count Technique and List Experiment` (with G. Blair)
  - Implements all methods mentioned so far and more

The Endorsement Experiment: An Example

- Select policies and ask respondents for their support level
- Randomly assign endorsers to respondents to endorsers
- No endorser for the control group
- The script for the control group:

```
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?

Strongly agree   Agree   Indifferent   Disagree
Strongly disagree   Don't Know   Refuse to answer
```

The Endorsement Experiment: An Example

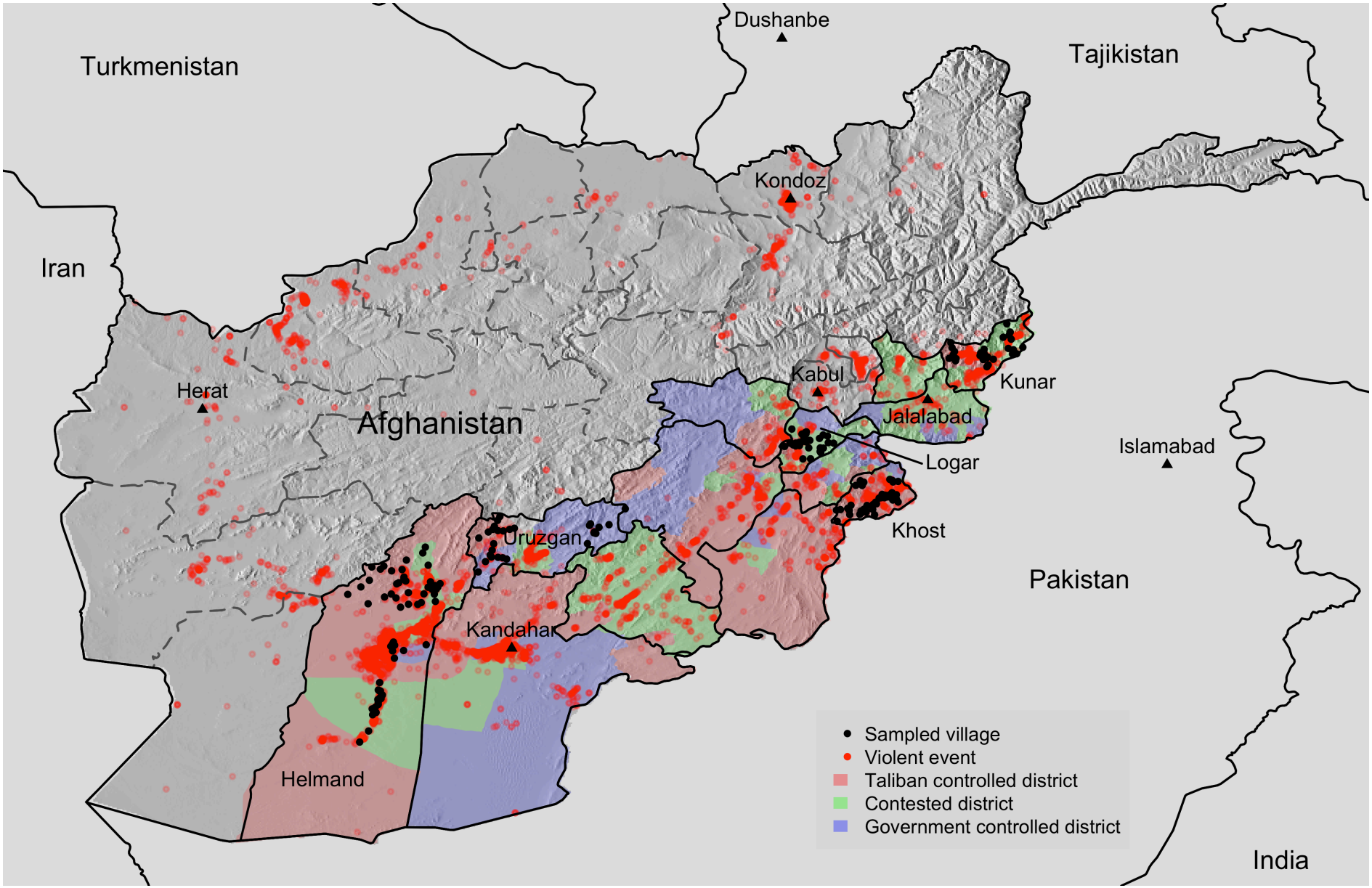
- Select policies and ask respondents for their support level
- Randomly assign endorsers to respondents to endorsers
- No endorser for the control group
- The script for a treatment group:

A recent proposal by the Taliban 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?

Strongly agree Agree Indifferent Disagree
Strongly disagree Don't Know Refuse to answer

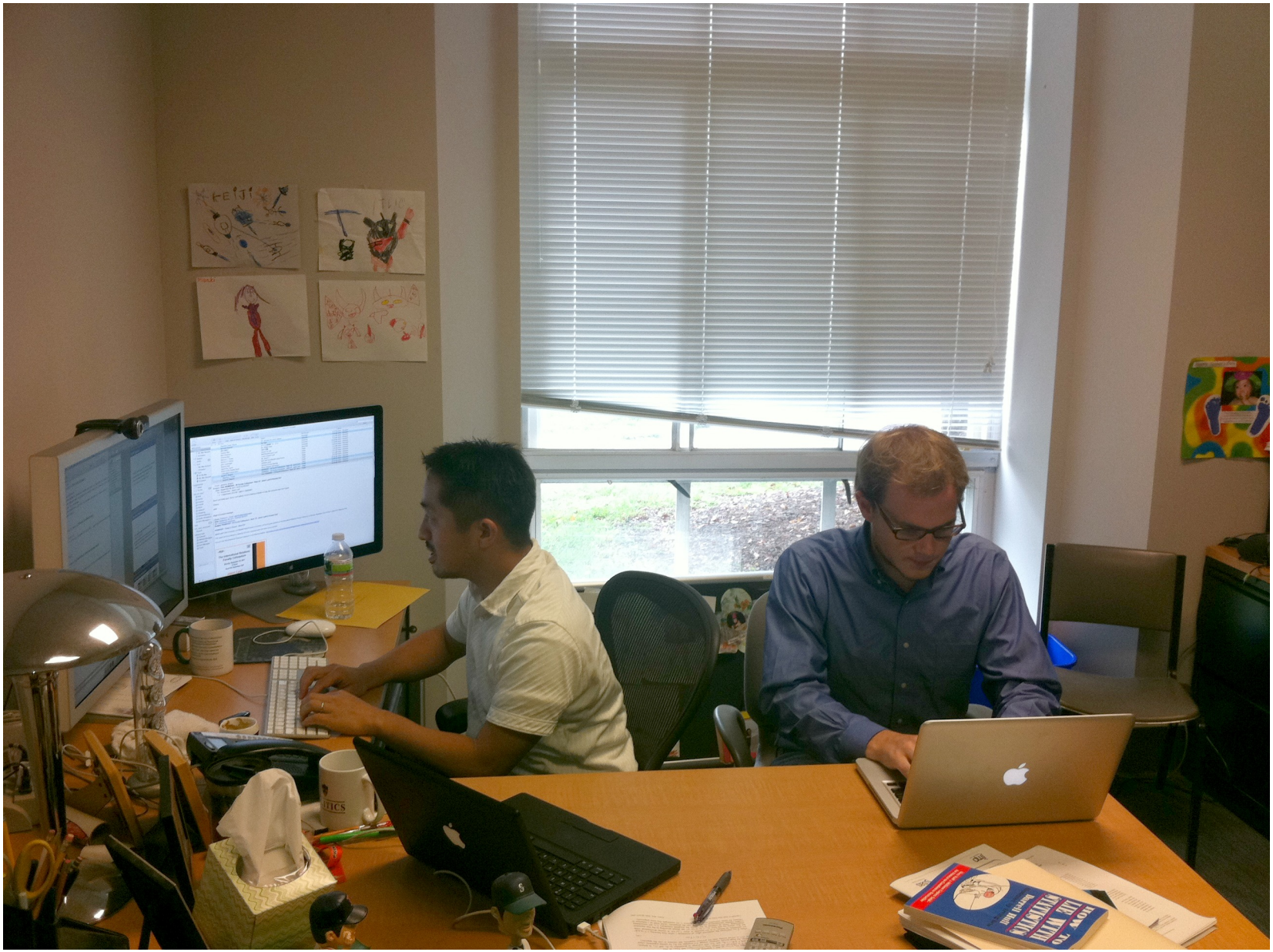
Public Nature of Interviews

Sampling in the Heartland of Insurgency



Negotiated Access

A Battlefield in Princeton, New Jersey

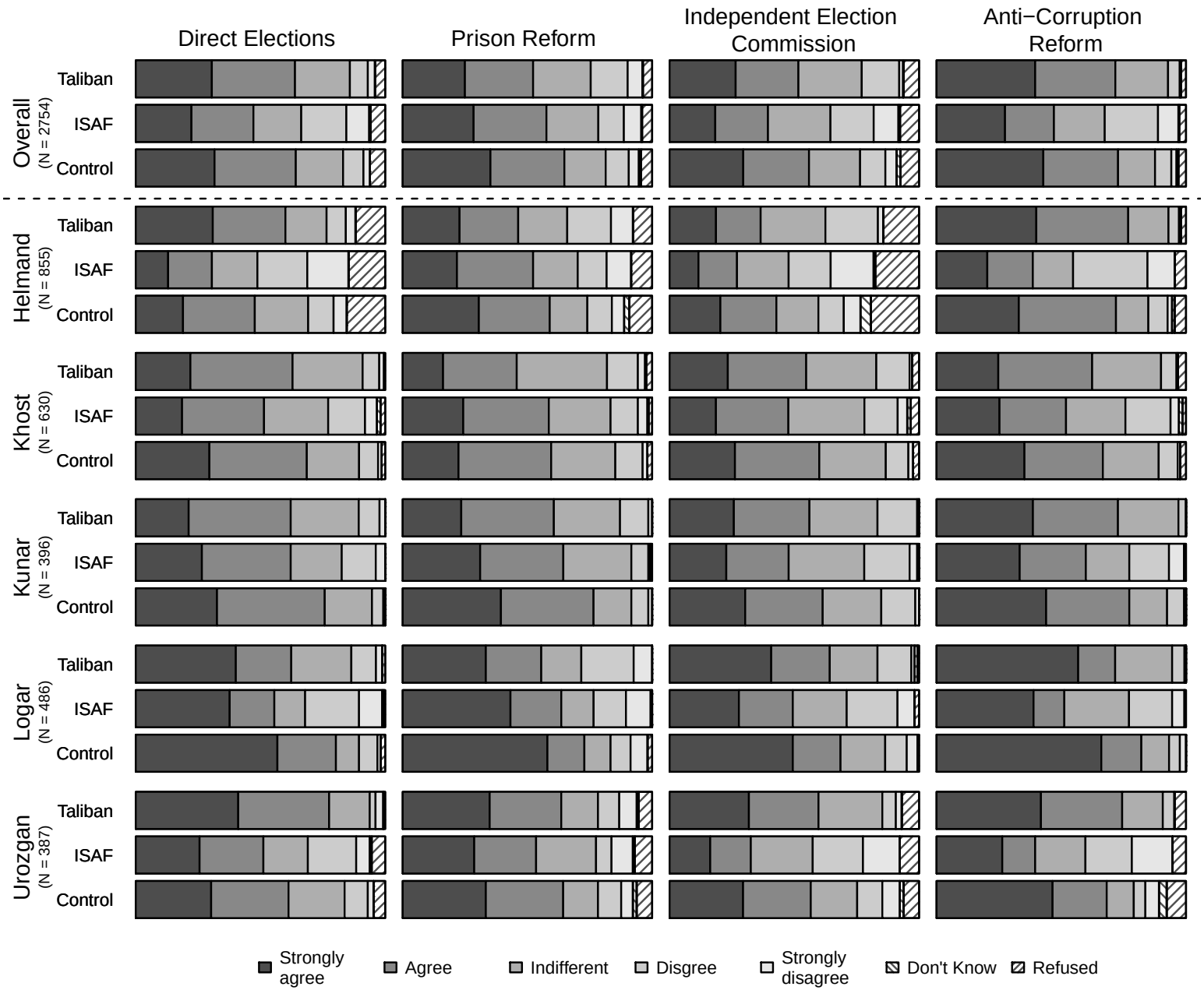


The Endorsement Experiment

- Carefully selected four "reform" policies:
  1. Direct elections
  2. Prison reform
  3. Independent election commission
  4. Anti-corruption reform

- Detailed justification for each policy: e.g., prison reform
  - Notoriously corrupt and inefficient
  - Similar proposals by the Taliban and ISAF
  - 48% of Afghans have no faith in prisons (Asia foundation)

Distribution of Responses



Endorsement Experiments Framework

- N respondents
- J policy questions
- K political actors
- $Y_{ij} \in \{0, 1\}$: response of respondent i to policy question j
- $T_{ij} \in \{0, 1, \ldots, K\}$: political actor randomly assigned to endorse policy j for respondent i
- For the Afghan experiment, an individual receives the same treatment across policies $T_i = T_{ij}$
- Covariates measured prior to the treatment

Statistical Modeling

- Multiple questions $\Longrightarrow$ item response theory

$$\Pr(Y_{ij} = 1 \mid T_i = k) \;=\; \Phi(\alpha_j + \beta_j(x_i + s^*_{ijk}))$$

  - α_j: average popularity of policy j
  - β_j: how much policy j differentiates pro- and anti-reform respondents
  - x_i: "ideal point" $=$ how pro-reform respondent i is
  - s_{ijk}: influence of endorsement by group k

- Support level:

$$\frac{\partial}{\partial s_{ijk}} \Pr(Y_{ij} = 1 \mid T_{ij} = k) > 0$$

where

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

The Proposed Model (Continued)

- Multi-stage sampling $\Longrightarrow$ Multi-level modeling

$$
\begin{aligned}
s_{ijk} &\overset{\text{indep.}}{\sim} \mathcal{N}(\lambda_{k,\text{village}[i]} + Z_i^\top \lambda_k^Z,\ \omega^2_{k,\text{village}}) \\
\lambda_{k,\text{village}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\lambda_{k,\text{district}[i]} + V_{\text{village}[i]}^\top \lambda_k^V,\ \omega^2_{k,\text{district}}) \\
\lambda_{k,\text{district}[i]} &\overset{\text{indep.}}{\sim} \mathcal{N}(\lambda_{k,\text{province}[i]} + W_{\text{district}[i]}^\top \lambda_k^W,\ \omega^2_{k,\text{province}})
\end{aligned}
$$

- Same hierarchical structure for ideal points x_i
- "Noninformative" hyper prior on $(\alpha_j, \beta_j, \delta, \theta_k, \omega^2_{jk}, \Phi_k)$

- Interpretation:
  - spacial model vs. factor analysis
  - learning vs. support

Quantities of Interest

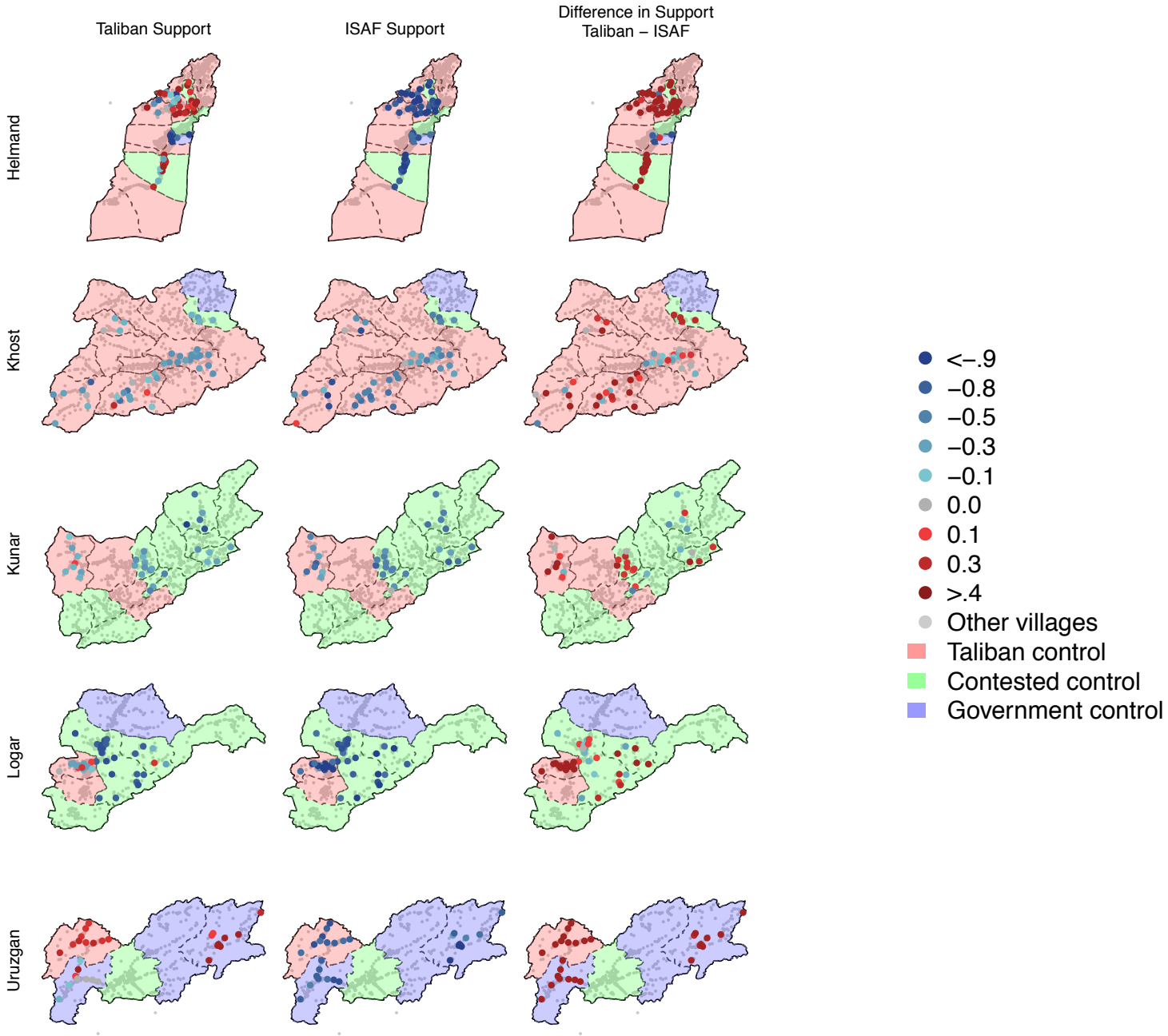
- Average support level for each group k

$$\begin{aligned} \tau_{jk}(Z_i) &= Z_i^\top \lambda_{jk} \quad \text{for each policy } j \\ \kappa_k(Z_i) &= Z_i^\top \theta_k \quad \text{averaging over all policies} \end{aligned}$$

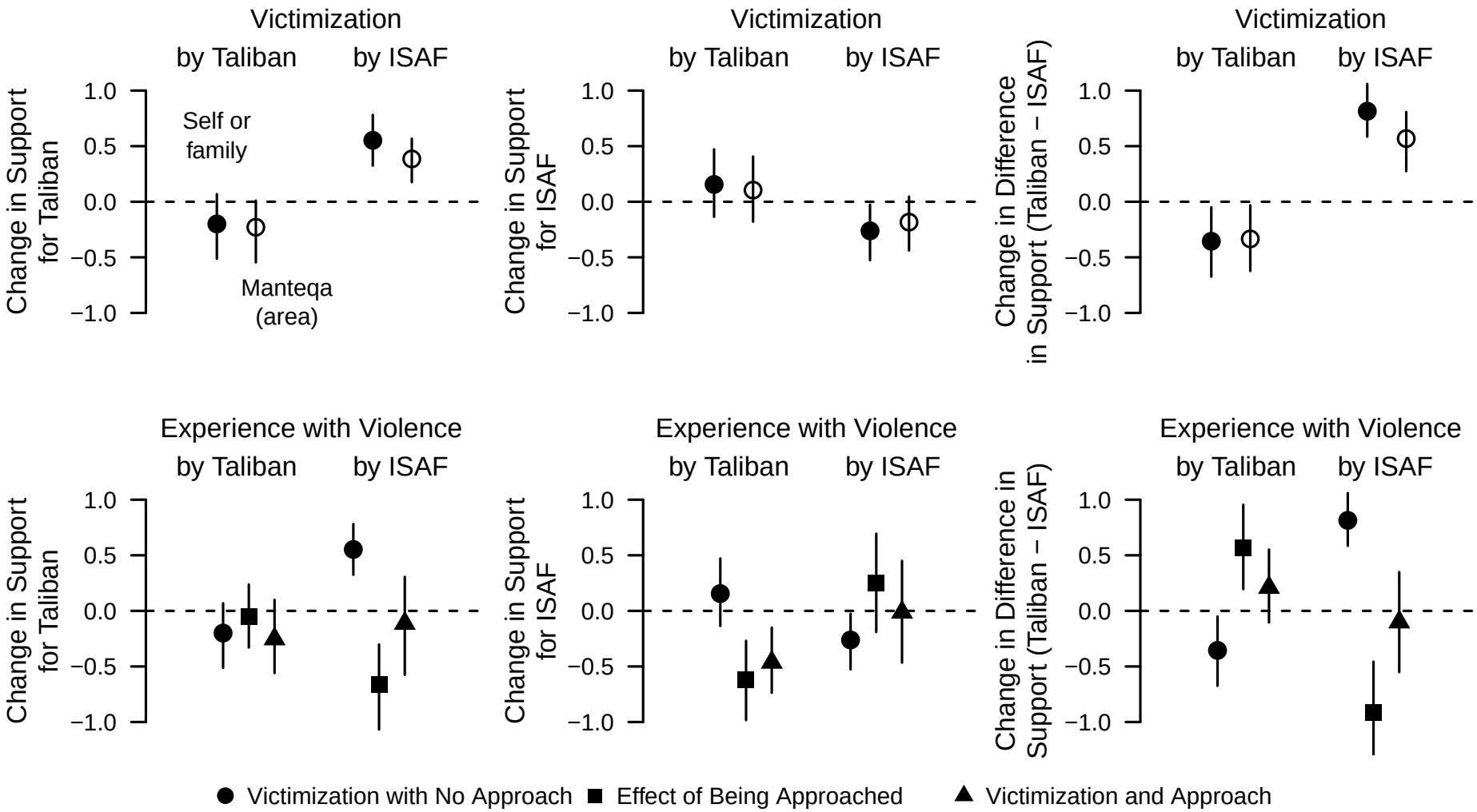
- Standardize them by dividing the (posterior) standard deviation of ideal points
- Proportion of supporters for each group k

$$\Pr(\kappa_k(Z_i) > 0)$$

Village Level Estimates

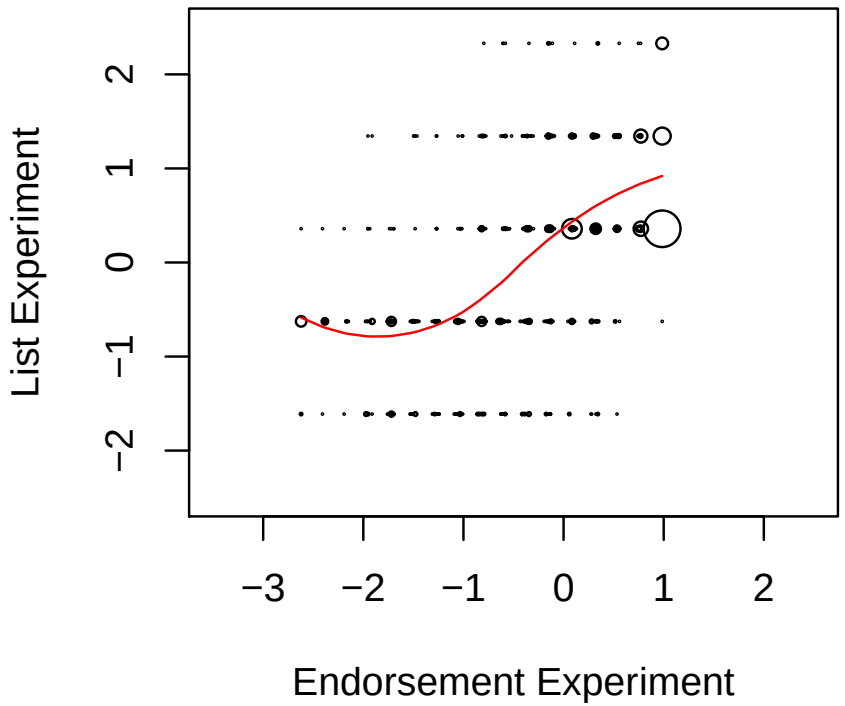


Effects of Victimization and Restitution

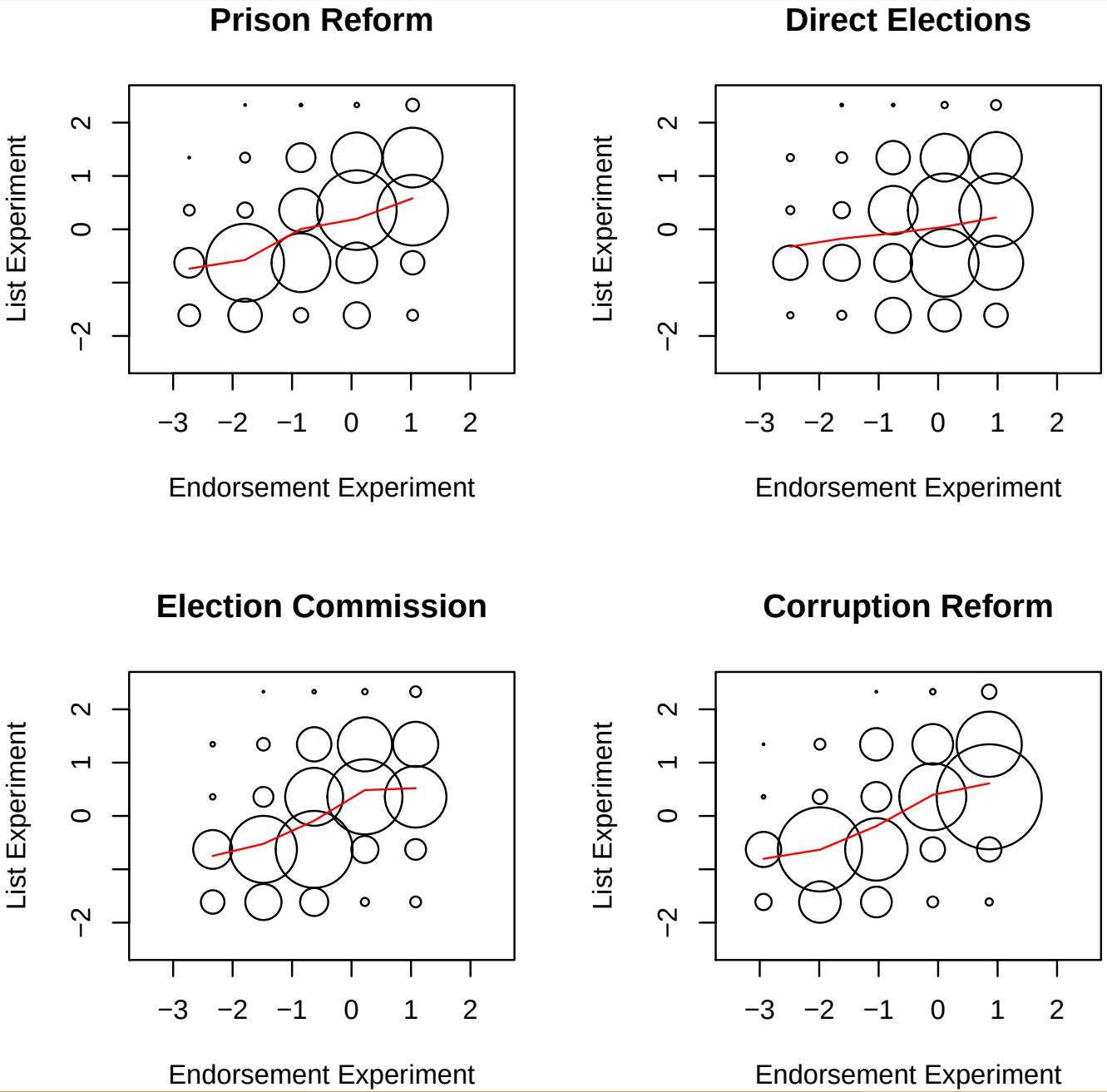


Comparison with List Experiments

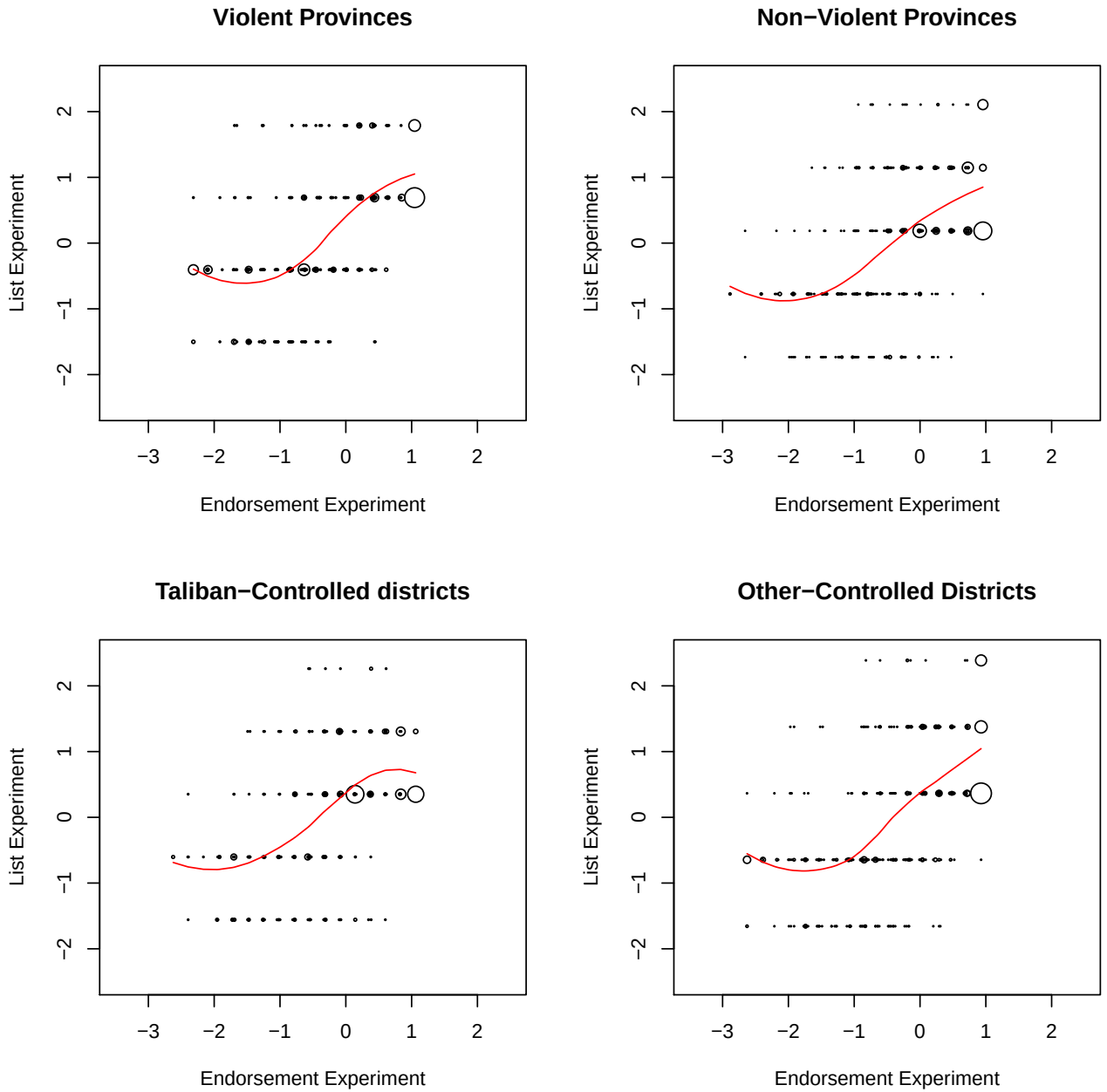
- Endorsement experiments seem to have worked quite well
- Challenges for list experiments in Afghanistan
  1. Illiteracy: median years of education $= 0$
  2. Public nature of interviews
- Result: *Massive* floor and ceiling effects for the Taliban list
- What about the ISAF list?



Comparison by Policy Questions



Comparison by Provinces and Districts



Practical Suggestions for Endorsement Experiments

- Analyzing endorsement experiments:
  - Pool several policy questions via item response theory
  - Multilevel modeling to efficient estimation of spatial patterns

- Designing endorsement experiments:
  - Policy positions should be well-known
  - Response distribution should not be skewed
  - Policies should belong to a single dimension

- Open source software:
  - `JAGS` (Plummer)
  - `endorse: R Package for Analyzing Endorsement Experiments` (with Y. Shiraito)

Concluding Remarks and Ongoing Research

- Challenges of eliciting truthful responses to sensitive questions:
  1. Bias reduction vs. Information loss
  2. Role of statistics $=$ efficient design and analysis
  3. Difficulty of validation $\Longrightarrow$ multiple measures, "ground-truthing"
  4. Art of survey research $+$ General methodological guidelines

- Ongoing research:
  1. Formally and empirically comparing and combining randomized response, and list and endorsement experiments
  2. Applications in various parts of the world (with collaborators):
     - Afghanistan
     - Columbia
     - Mexico
     - Nigeria
     - Pakistan
     - United States

The project website for papers and software:

http://imai.princeton.edu/projects/sensitive.html

Email for comments and suggestions:

kimai@Princeton.Edu