Eliciting Truthful Responses to Sensitive Survey Questions Using List Experiments

Kosuke Imai

Princeton University

Talk at the Inter-American Development Bank
August 17, 2017

Motivation

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

- Sensitive questions $\rightsquigarrow$ social desirability, privacy concerns
- Prejudice, corruption, drug use, sexual behavior, etc.
- Lies and nonresponses $\rightsquigarrow$ potential bias

- Survey experiments as a solution:
  1. aggregation $\rightsquigarrow$ List experiment (item count technique)
  2. randomization $\rightsquigarrow$ Randomized response method
  3. cueing $\rightsquigarrow$ Endorsement experiment

Overview of the Talk

- Methodological developments for list experiments:
  1. multivariate regression for list experiments (Imai, 2011; Imai, Park & Greene 2015)
  2. statistical test for violation of assumptions (Blair and Imai, 2012)
  3. modeling ceiling and floor effects (Blair and Imai, 2012)

- Improving list experiments:
  1. comparing and combining with other methods (Blair, Imai & Lyall 2014)
  2. incorporating auxiliary information (Chou, Imai, & Rosenfeld, Forthcoming)

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.

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
- The EM algorithm for reliable estimation
- The ML estimator is more efficient but less robust than the NLS estimator
- Both NLS and ML estimators can be extended to use the latent response to a sensitive item as an explanatory variable in a regression

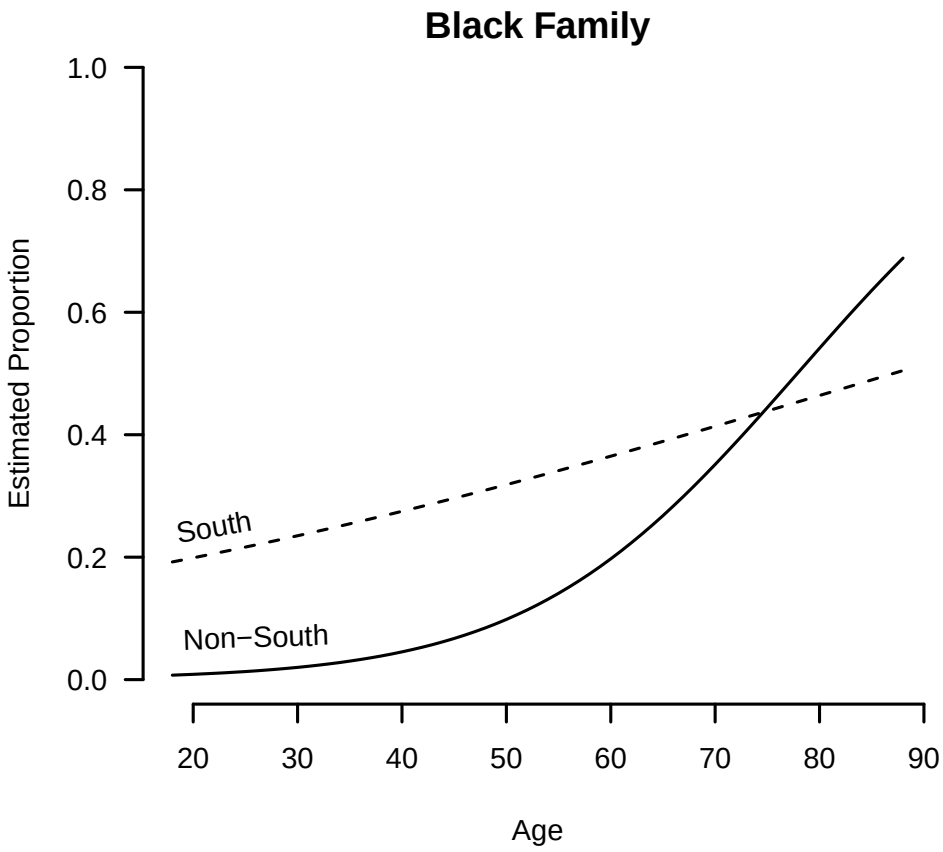
Empirical Application: Racial Prejudice in the US

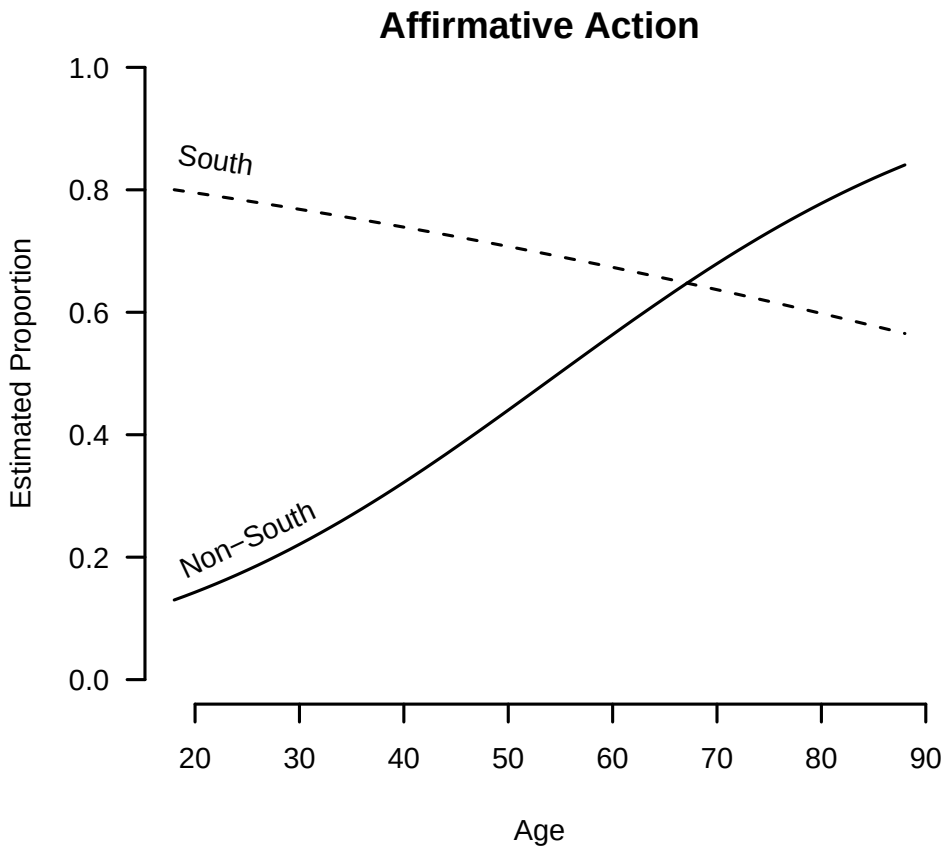
- Kuklinski *et al.* (1997 JOP): Southern whites are more prejudiced against blacks than non-southern whites – no "New South"
- The limitation of the original analysis:

  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

Generational Changes in South and Non-South





- Age is important even after controlling for gender and education

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
  4. Consider alternative designs such as double list experiment

- Open-source software:
  - R package **list:** Statistical Methods for the Item Count Technique and List Experiment
  - Implements all methods mentioned so far and more

Two Ways to Improve Sensitive Question Survey Methods

1. Comparing and combining multiple measurements (Blair, Imai & Lyall 2014)
   - Agreement among multiple measurements $\rightsquigarrow$ more credible
   - Combining multiple measurements $\rightsquigarrow$ more powerful
   - Application: Hearts and minds in Afghanistan

2. Using auxiliary information (Chou, Imai, & Rosenfeld Forthcoming)
   - Sometimes aggregate truths are available
     - Turnout rates and voting outcomes
     - Administrative records, e.g., crime and incarceration
   - Use auxiliary information to improve individual-level inference
   - Application: Missisippi anti-abortion referendum

Hearts and Minds in Afghanistan (Blair, Imai & Lyall, 2014)

- How do we measure civilian attitudes in a conflict setting?
- Current efforts in Afghanistan rely on direct questions:
  1. USAID (TCAPF): "Who do you believe can solve your problems?"
  2. ISAF (ANQAR): "Over the past 6 months, do you think the Taliban have grown stronger, grown weaker, or remained the same?"
- Why are direct questions a bad idea?
  1. Threats to enumerators and respondents
  2. Nonresponse, social desirability bias
  3. Interviews are public
  4. Danger of selection bias in sampling locations (role of gatekeepers)
- ANQAR (November-December 2011): 50% refusal rate

Public Nature of Interviews

List Experiments

- Script for 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

List Experiments

- Script for the treatment 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; **ISAF**

The Data from the List Experiment

response value	Control Group frequency	Control Group proportion	ISAF Treatment Group frequency	ISAF Treatment Group 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	

Endorsement Experiments

- 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

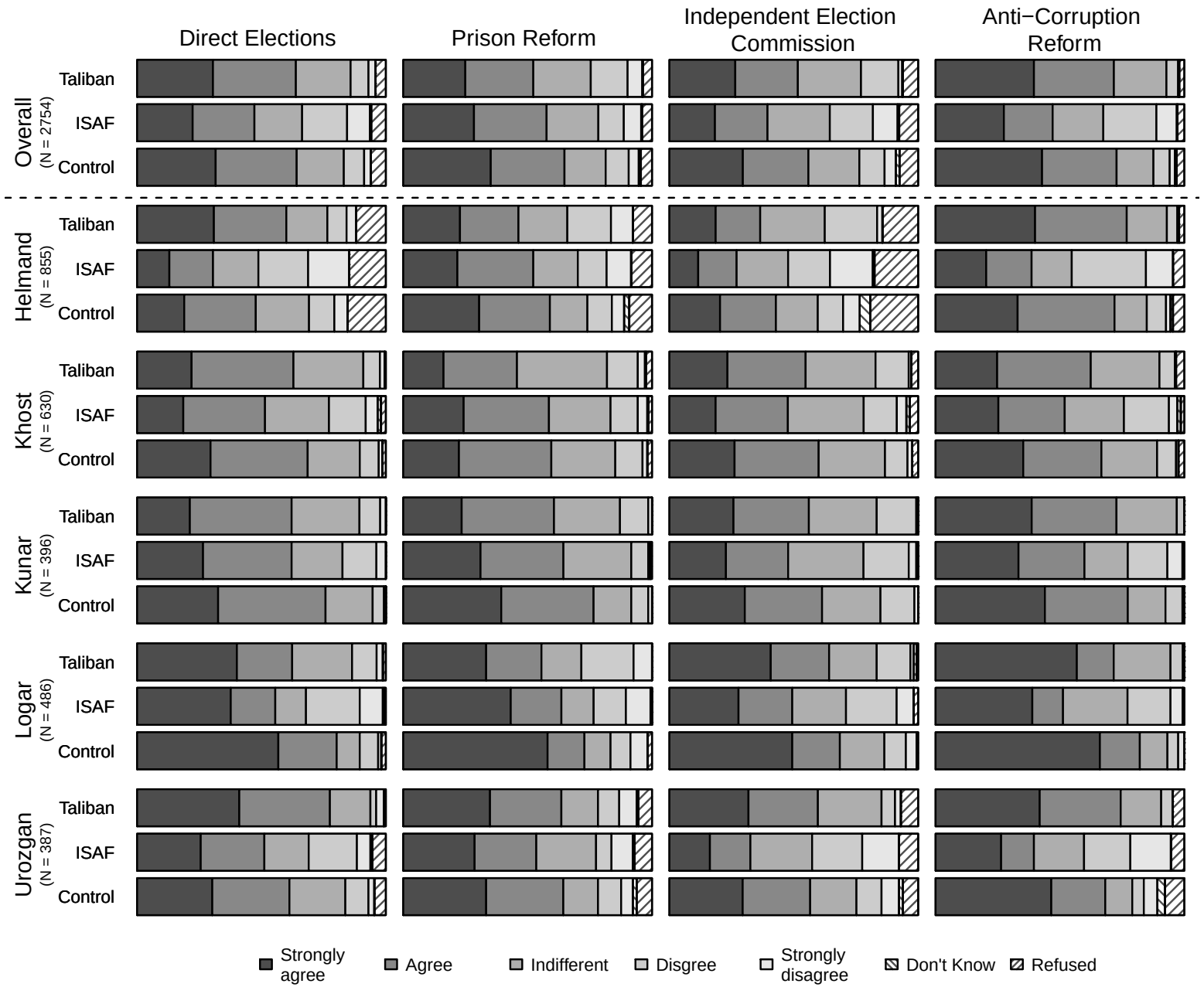
Endorsement Experiments

- Script for the treatment group:

  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?

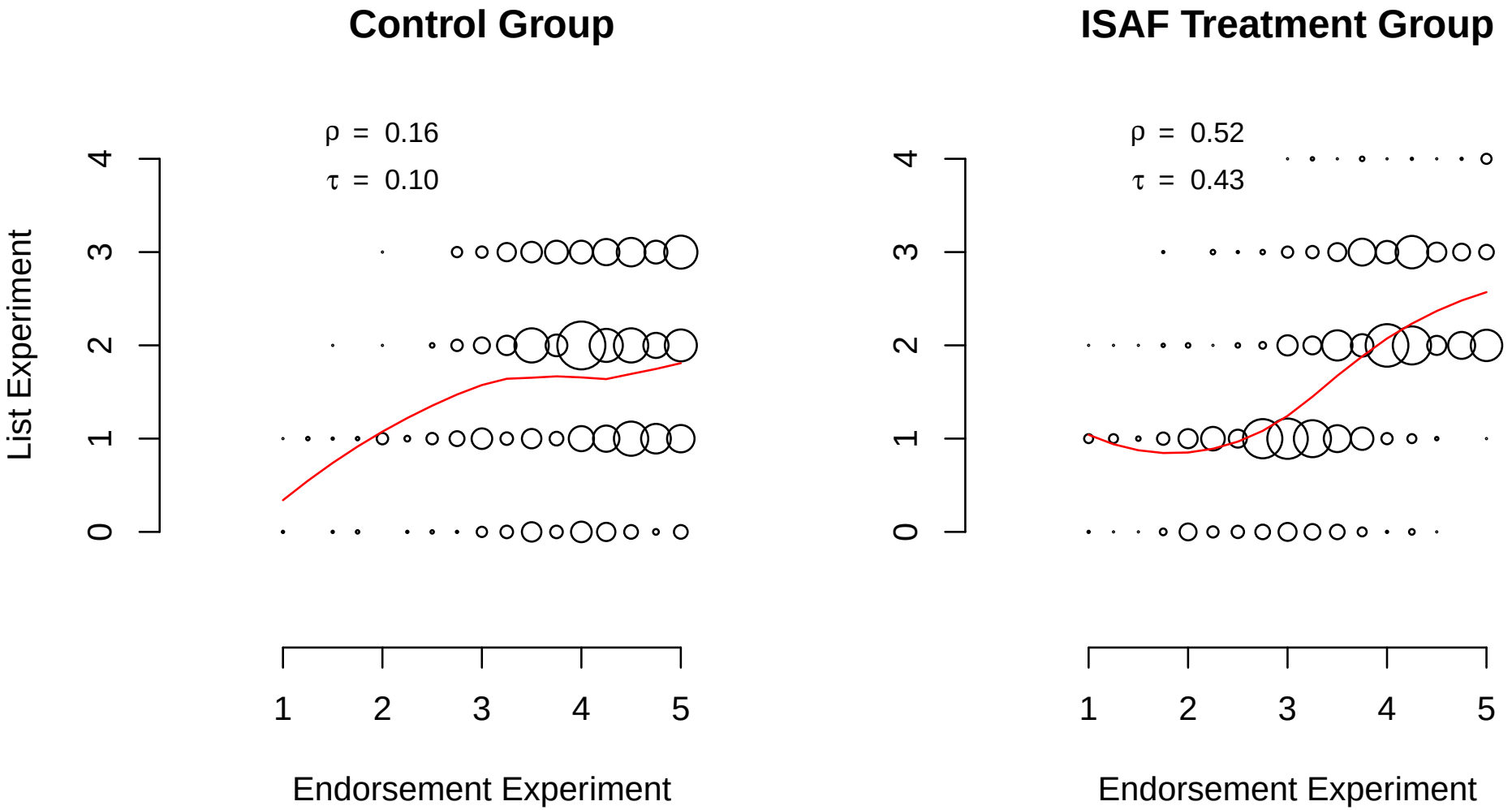
  Strongly agree; Agree; Indifferent; Disagree; Strongly disagree; Don't Know; Refuse to answer

Data from the Endorsement Experiments



Descriptive Comparison: Overall

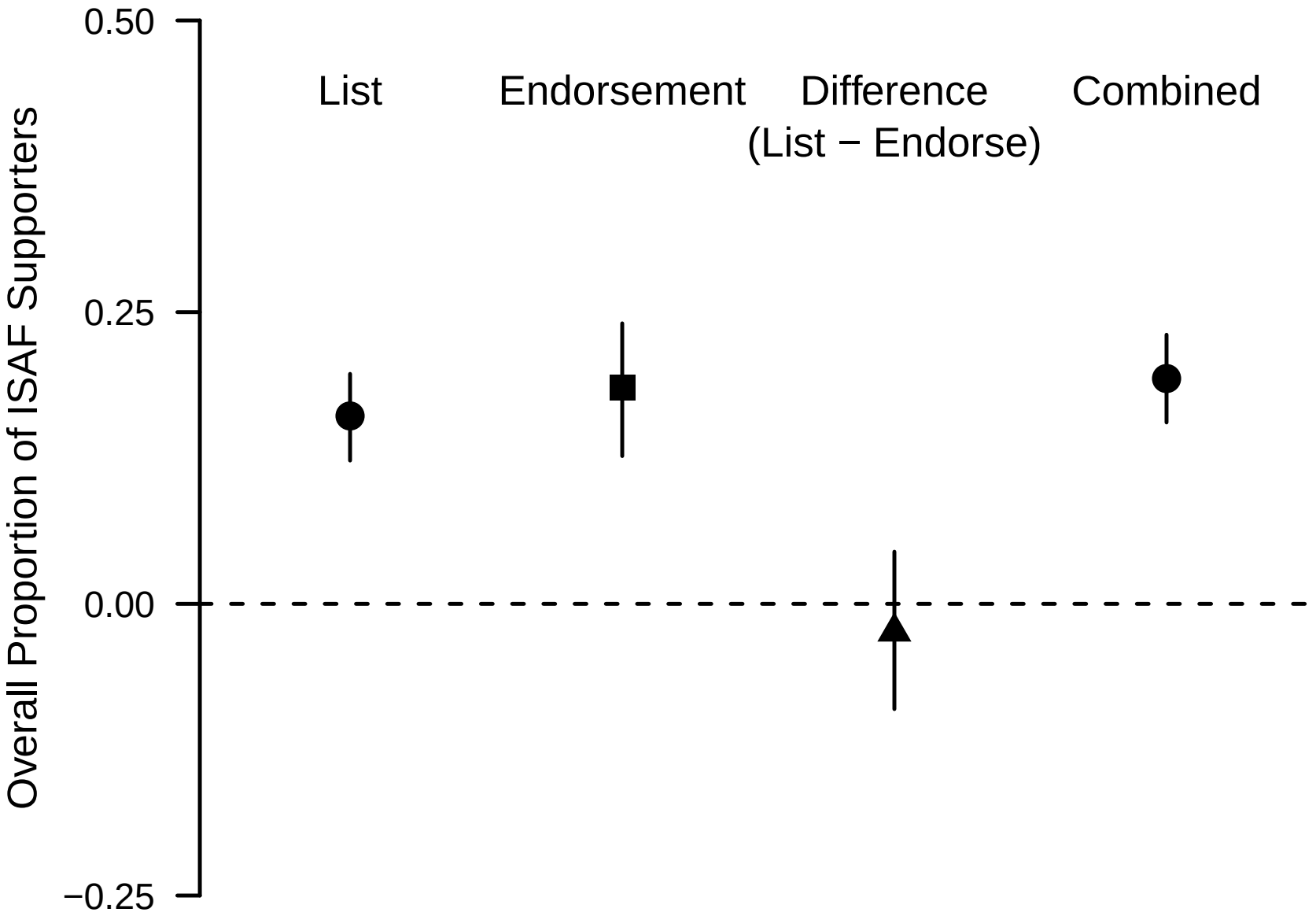
- Need for validation $\rightsquigarrow$ Multiple measurement strategy
- Two measures should give similar results



Comparing and Combining List and Endorse Experiments

- Formal comparison and integration
- What is the probability of supporting ISAF?
  1. List: prob. of saying yes to the sensitive item
  2. Endorsement: prob. of endorsement having a positive effect on support for policy
- These probabilities should be similar!
- They can be estimated with a new multivariate regression method
- Endorsement and list experiments can even be combined for a joint analysis

Overall Proportion of ISAF Supporters



The Mississippi Validation Study (Rosenfeld, Imai & Shapiro 2016)

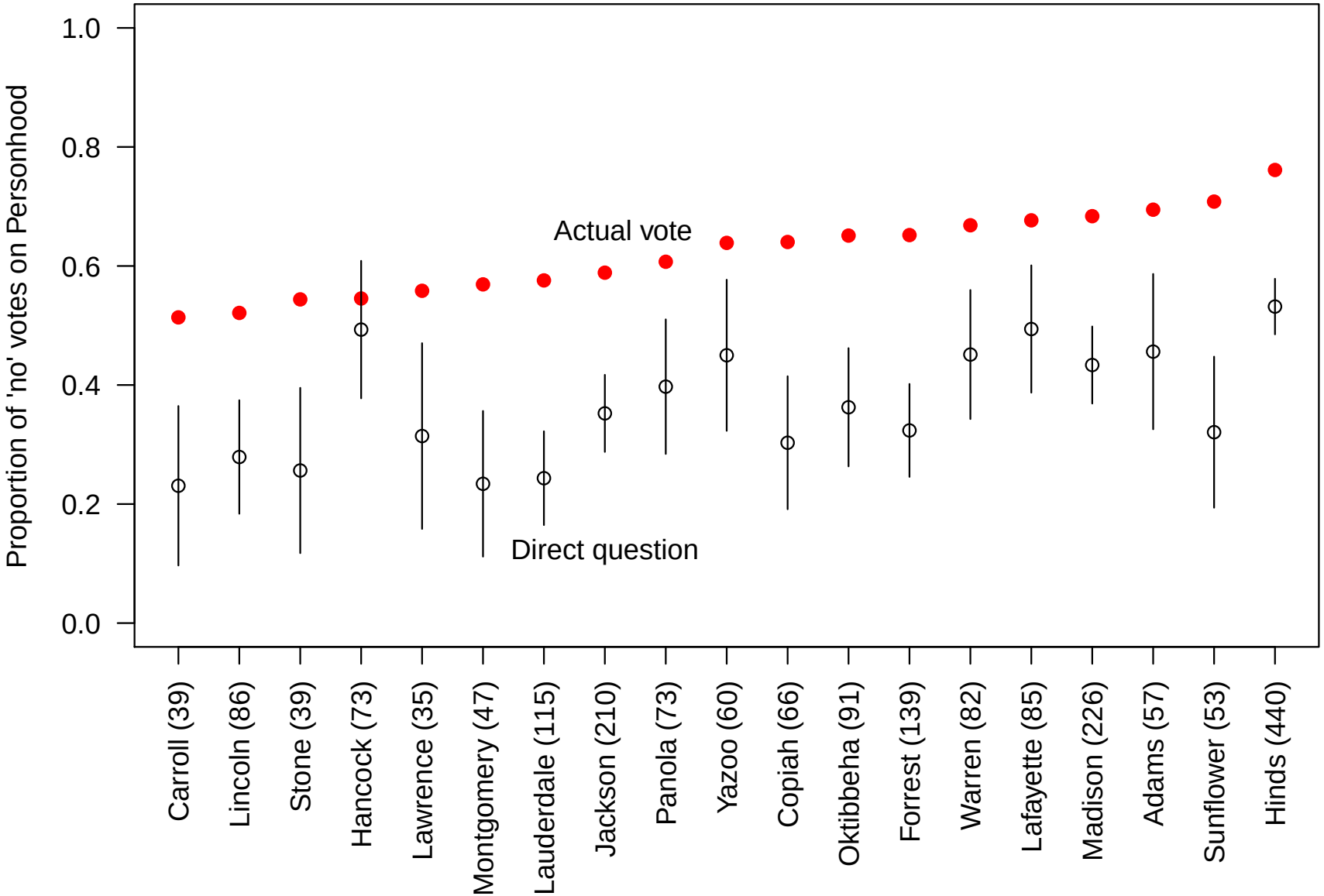
- Estimate voting on anti-abortion referendum using:
  - direct question
  - list experiment (item/unmatched count technique)
  - endorsement experiment
  - randomized response
- Validate estimates against official election outcome:
  - sample from voter history file
  - county-level voting recap reports for validation
- Case selection:
  - a poll conducted 24 hours before the election predicts 44% no votes
  - the amendment was defeated 58% to 42%
- Findings:
  - direct question $\rightsquigarrow$ significant under-estimation though efficient
  - indirect methods $\rightsquigarrow$ much less biased though more variable
  - endorsement and randomized response $\rightsquigarrow$ least bias

Direct Question

Did you vote YES or NO on the Personhood Initiative, which appeared on the November 2011 Mississippi General Election ballot?

- Voted Yes
- Voted No
- Did not vote
- Don't know
- Refused

Bias of the Direct Question



List Experiment

Here is a list of four things that some people have done and some people have not. Please listen to them and then tell me HOW MANY of them you have done in the past two years. Do not tell me which you have and have not done. Just tell me how many:

Discussed politics with family or friends

Cast a ballot for Governor Phil Bryant

Paid dues to a union

Given money to a Tea Party candidate or organization

(treatment) Voted 'YES' on the 'Personhood' Initiative

How many of these things have you done in the past two years?

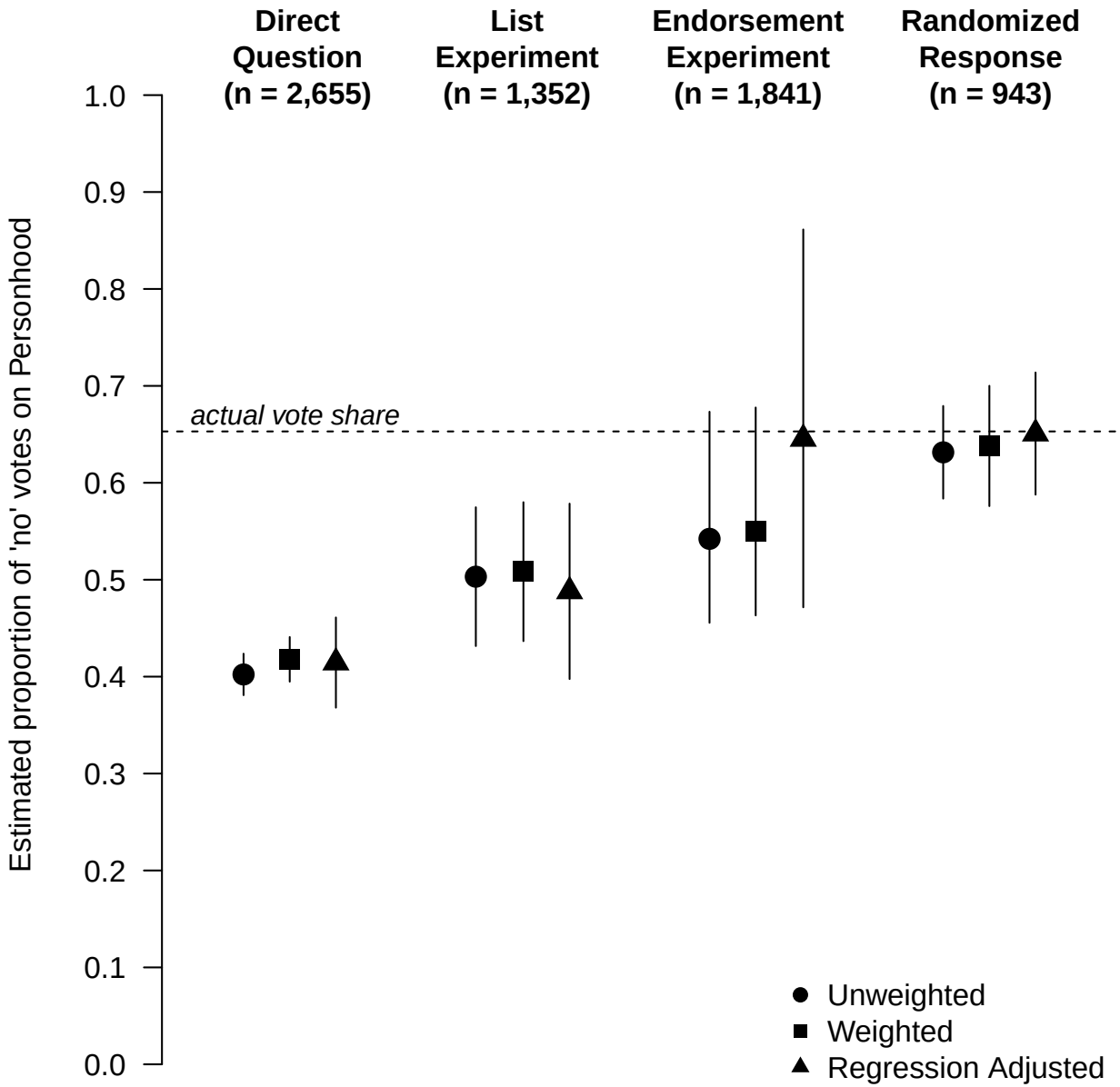
Endorsement Experiment

We'd like to get your overall opinion of some people in the news. As I read each name, please say if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of each person.

(control) Phil Bryant, Governor of Mississippi?

(treatment) Phil Bryant, Governor of Mississippi, who campaigned in favor of the 'Personhood' Initiative on the 2011 Mississippi General Election ballot?

Pooled Analysis



The Proposed Methodology for List Experiment

- List experiment can be analyzed by method of moments:

$$\mathbb{E}(Y_i \mid T_i, X_i) = \underbrace{f(X_i, \gamma)}_{\text{Control Items}} + T_i \underbrace{g(X_i, \delta)}_{\text{Sensitive Trait}}$$

- We simply add moment conditions of the form

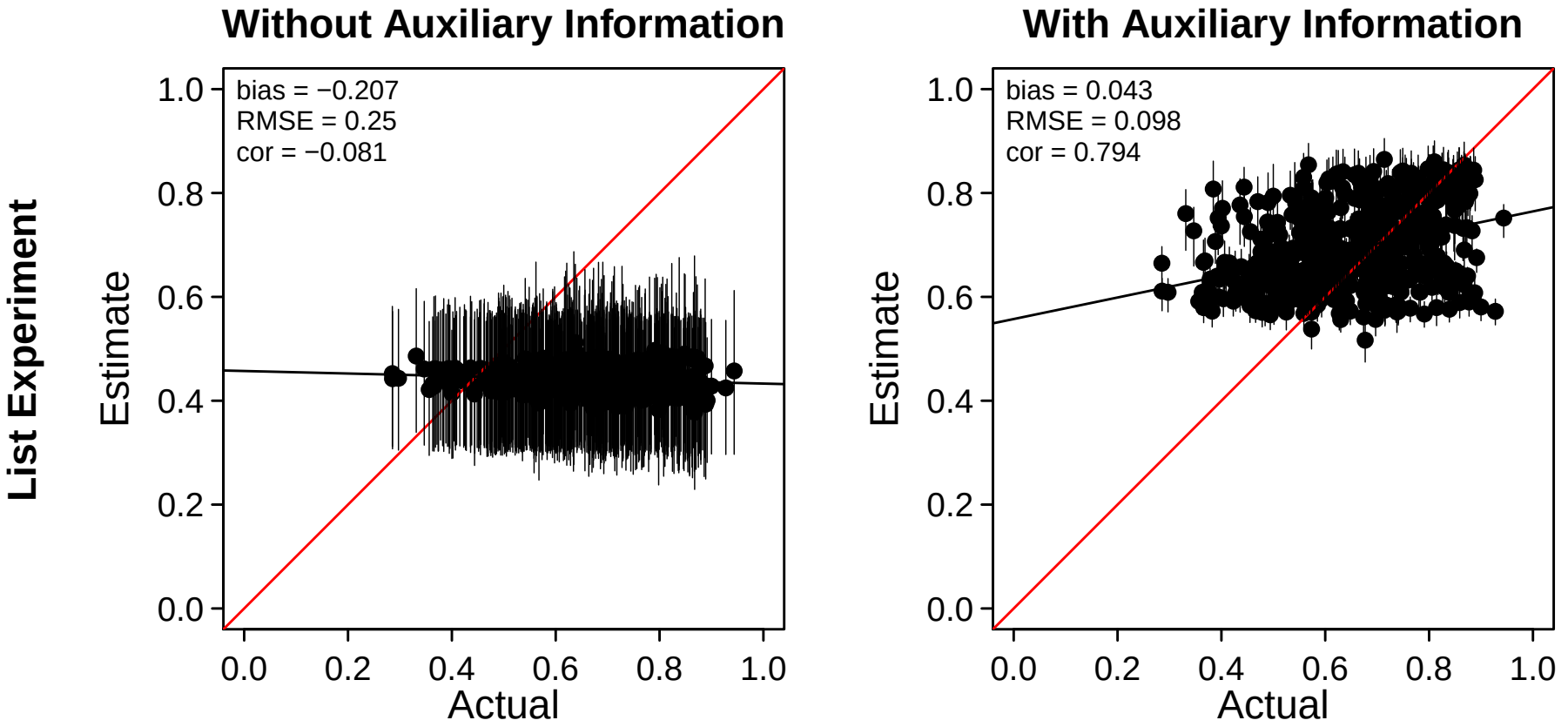
$$\begin{aligned} \mathbb{E}[g(X_i, \delta)] &= h \\ \mathbb{E}[g(X_i, \delta) \mid G_i = k] &= h_k \end{aligned}$$

- A similar strategy works for randomized experiment
- (Testable) Assumption: Same parameters solve all moment conditions
  $\rightsquigarrow$ Constant parameters across groups.

Efficiency Comparison with Direct Questioning

	List Experiment N=1,325		**Randomized Response** N=818		**Endorsement Experiment** N=1,841	
	s.e.	ratio	s.e.	ratio	s.e.	ratio
Direct questioning	0.017		0.021		0.289	
No auxiliary info.	0.067	3.963	0.040	1.928	0.348	1.204
With auxiliary info.	0.019	1.150	0.018	0.855	0.276	0.955

Auxiliary Information Improves List Experiment



- Improves endorsement experiment and randomized response method
- Improve multivariate regression estimates

Concluding Remarks

- Direct question is often severely biased
- All indirect methods can reduce bias:
  - Degree of privacy protection: endorse $>$ randomized response $>$ list
  - Ease of implementation: list $>$ endorse $>$ randomized response
- But, they are inefficient: bias-variance tradeoff
- Two ways to improve indirect question methods:
  1. Use of multiple-measurement strategies when truth is not available
  2. Use aggregate-level truth to improve individual-level estimates
- Open-source software:
  - `list` for list experiment
  - `endorse` for endorsement experiment
  - `rr` for randomized response

The project website for papers and software:

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

Email for comments and suggestions:

kimai@Princeton.Edu