Appendix for Means, Motive, & Opportunity in Becoming Informed About Politics: A Deliberative Field Experiment with Members of Congress and Their Constituents

Kevin M. Esterling
(Corresponding Author)
Associate Professor
Department of Political Science
UC–Riverside
900 University Ave.
Riverside, CA 92506
Tel. 951-827-3833
Fax 951-827-3933
kevin.esterling@ucr.edu

Michael A. Neblo
Assistant Professor
Department of Political Science
Ohio State University
2114 Derby Hall
154 N. Oval Mall
Columbus, Ohio 43210
Tel. 614-292-7839
Fax 614-292-1146
neblo.1@osu.edu

David M.J. Lazer
Associate Professor
Political Science & Computer Science
Northeastern University
301 Meserve Hall
Boston, MA 02115
Tel. 617-373-2796
Fax 617-373-5311
davelazer@gmail.com

A Introduction

This appendix describes the experimental design and statistical analysis for the paper "Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment with Members of Congress and their Constituents."

B Experimental Design

The experimental design and data collection for this field experiment are summarized in the paper. Here we describe the experimental design and the data we collected in more detail. In this section, we also discuss some deviations from the ideal experimental design – the kinds of complications that can often occur in a large field experiment – and how our methods address these issues. In all cases where the methodological issue centers on missing data, we report sensitivity analyses to assess the outer limits of how our results conceivably could change, under extreme assumptions of what we could have observed had the data not been missing.

B.1 Subject Recruitment and Selection

Figure A1 gives an overview of the assignment, compliance and response rates among the 2222 subjects in the experiment. The flow chart has five stages: an RSVP, assignment, and then (if eligible) exposure to the background materials (BGM), exposure to a deliberative session, and a follow up survey.

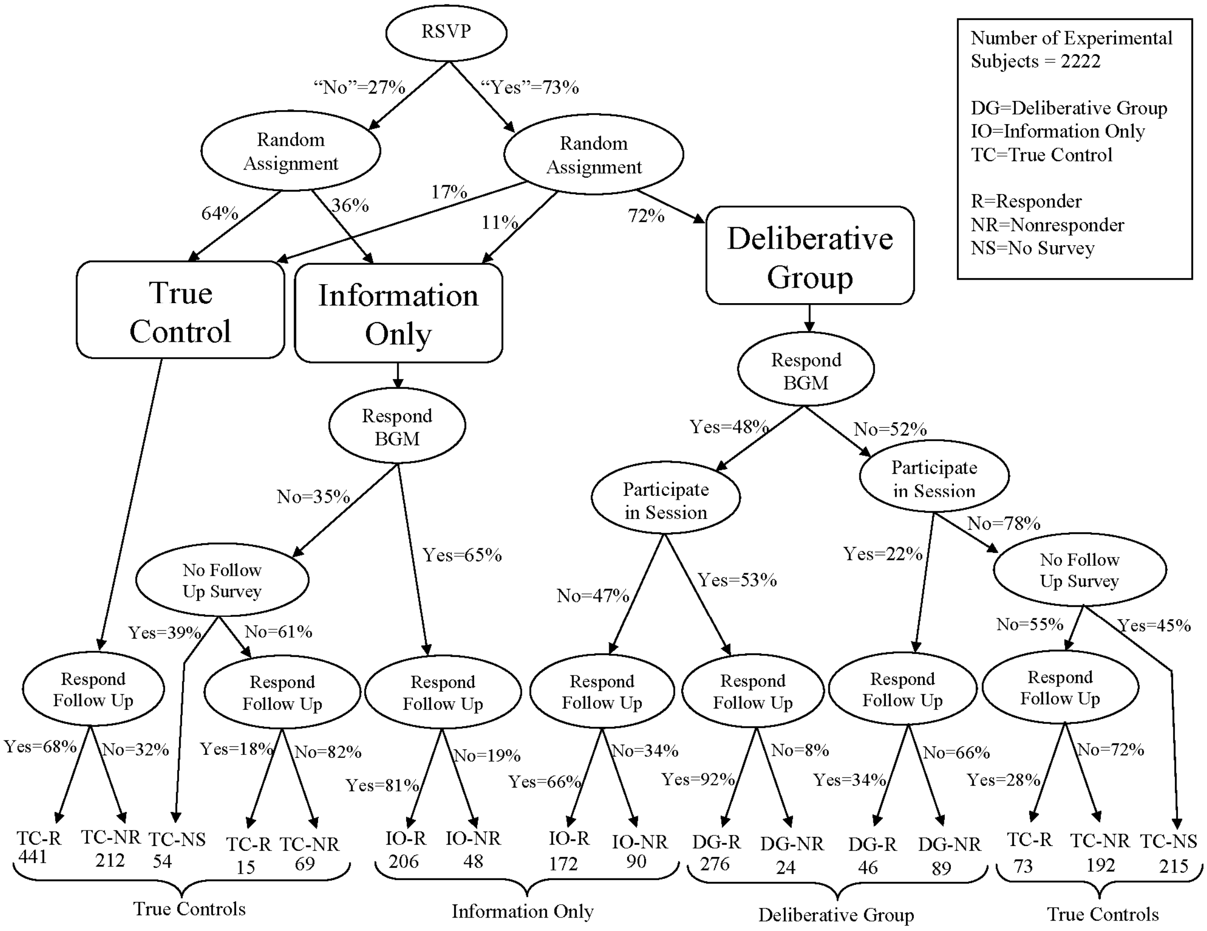

Figure A1: Assignment, Compliance and Response Rates

B.1.1 RSVP and Assignment

In the baseline survey we included an RSVP filter question, which indicated the time the session would take place for the subject's congressional district, and that the session would last approximately an hour. We then allowed subjects to indicate whether they 1) would be willing and able to attend the session; 2) would only complete surveys for the project; and 3) refused to participate in the project. Only 10.7 percent of respondents refused to participate in the study; we discard these observations and do not consider them further.[1] Among those who agreed to participate in some way, 73 percent of subjects indicated they would be willing and able to attend a session, and these subjects were randomized among the three treatment arms (72 percent to deliberative group (DG) condition, 11 percent to the information only (IO) condition, and 17 percent to the true control (TC) condition). The remaining 27 percent of subjects who indicated they wanted to participate in the study but would not attend the session were randomized among the information only condition (36 percent) and the true control condition (64 percent).

We chose these assignment rates in an attempt to create as-treated groups of sufficient size, for each treatment. For example, we assigned relatively few subjects to the IO condition, since we assumed many subjects assigned to the DG condition would read and complete the background materials (BGM) survey but fail to attend their assigned session. These subjects would then receive the IO treatment. Of course, we did not know the compliance rates in advance, so the as-treated cell sizes are not identical to each other.

We included the RSVP filter question to improve the information we had available at the initial assignment stage. When we began the study, KN did not know the rate at which subjects would attend the session in practice. As a result, at the beginning of the study, we simply did not know what proportion of subjects to assign to the deliberative condition in order to ensure enough subjects who complete the deliberative session and who respond to the follow up survey. Asking the RSVP filter question gave us some of this information. And of course, once we asked the RSVP question, it would have been odd to invite subjects after they

[1] An additional 299 subjects did not respond to the baseline survey, for an AAPOR RR6 response rate of 0.76 (see Callegaro and Disogra, 2008).

have indicated they would not attend. Hence, we randomize these participants among the other two arms.

We retain the participants who RSVP'ed "No" (self-reported they wished to participate in the study but would not attend a session) for use in the statistical analysis. Asking the RSVP question simply gave us more information regarding the likely take up rate of the deliberative sessions, but otherwise does not affect our analyses, under one assumption: that the RSVP self-reports are true and the respondent indeed would not have attended a session had she been given the opportunity.[2] To see this, imagine the case if we had not asked this filter question. Assuming those who RSVP'ed "No" would have failed to attend a session if invited, these respondents would have selected themselves into one of the control groups through noncompliance. As we emphasize in the paper, noncompliance is inevitable in a field experiment and we use statistical methods to identify causal effects in the presence of noncompliance. Thus, under the assumption that the RSVP "No's" are accurate, asking the RSVP gave us information for basing assignment rates, but otherwise is irrelevant to the study.[3]

Because RSVP "No" subjects were not invited to a deliberative session, we cannot know whether they in fact would have attended if given the chance, and what their responses might have been had they attended. To address this limitation of our design, below we report sensitivity tests for how our treatment effect estimates would change under extreme assumptions regarding their compliance and regarding their responses to the post-treatment knowledge items. We find causal effects even under these extreme scenarios. Considering that

[2] In our statistical model, the converse of this assumption, that those who say they will attend in fact attend, does not need to be true.

the assumptions in the sensitivity test regarding compliance and treatment effects are themselves quite implausible, we are confident that our decision to include the RSVP filter has no consequences for our findings.

B.1.2 Treatment Compliance

The nodes below the assignments indicate compliance with each task (excluding the November survey, which we use only as an indicator of compliance type). Cell sizes for the treatment actually received (the cell "Ns") are indicated in the terminal nodes (the bottom row) of the diagram. The cell labels have two components. For the first component, DG (deliberative group) indicates the subject participated in a deliberative group; IO (information only) indicates the subject completed the informational background materials survey but not a session and hence received the information only treatment; and TC (true control) indicates the subject attended neither a session nor read the BGM material and hence is a true control subject. For the second component, R (responder) indicates the subject responded on the follow up survey; NR (nonresponder) indicates the subject was offered the follow up survey but chose not to respond; and NS (no survey) indicates the subject was not administered the follow up survey.

B.1.3 Administration of the Follow Up Survey

About halfway through the study, in negotiations over unexpectedly high costs with our survey vendor, we agreed to discontinue sending follow up surveys to subjects with the strongest histories of nonresponse. These were the subjects who, *ex ante*, were least likely to

[3] As we note in appendix B.3.2, opting out at the RSVP stage was statistically unrelated to compliance, so we do not use this self-report as a behavioral indicator in the compliance model.

respond to the follow up survey and thus were most likely to need to have their responses imputed anyway. As a result, a total of 269 subjects, or about 12 percent of the sample, did not receive a follow up survey. But note that even if we had sent these subjects the survey, most of them would not have filled them out. We know this because the revision to the survey procedures occurred after we had fielded the study to more than half of the sample. As a result, 349 of the subjects we later identified as "chronic nonresponders" were sent a follow up survey and, among these, only 25.7 percent responded. Thus, among the 269 who were not sent the survey (assuming that the order of the districts does not matter), we likely would have observed only an additional 67 surveys returned, which is only 3 percent of the total sample.

In principle, this revision to the survey procedures poses little problem for our statistical analysis. In the compliance model, response to the follow up survey is missing for these subjects; we do not treat these "nonresponses" as behavioral data, in that the statistical model does not assume this "failure" to respond to the follow up survey reveals any additional information at all about these subjects' compliance type. Instead, the model imputes their probability distribution of compliance based on their observed behavioral data and covariates. Likewise, the model imputes the probability distributions of their responses to the policy knowledge items on the follow up surveys based on their pretreatment knowledge, their compliance type, and important covariates. The model below accommodates these imputations by allowing our uncertainty about whether and how the respondent would have responded to propagate through all estimated parameters of the statistical model (Tanner and Wong, 1987).

As we do with the RSVP "No" respondents, in appendix B.3.2 we report sensitivity tests to assess how our treatment effect estimates would change under extreme assumptions of a pre-post knowledge change these respondents could have shown, even having not attended a session. We find that the results do not change whatsoever under even extreme assumptions of knowledge gains from this subset of respondents, including under the scenario where we assume all 269 of the chronic nonresponders responded and that each had an unusually large knowledge gain for a control subject. So again we are confident that excluding this subgroup of chronic nonresponders from the follow up survey has little or no consequences for our reported findings.

We only administered the November survey to participants who completed a follow up survey, and/or who participated in a deliberative session. We do not use responses to the November survey questions in any way in this analysis. Instead, we only use an indicator of whether or not participants returned this survey as an additional behavioral compliance indicator. For those who are not administered a November survey, we impute a probability distribution for their response based on pretreatment data and the latent compliance variable, just as we do for any other missing compliance indicators, such as for compliance with the treatment for those assigned to the control. The logic of this restriction, the imputation of their response if missing, and the consequences for estimation are identical to that of the follow up survey.

B.2 Background Reading Materials (BGM)

As we note in the paper, participants in the deliberative group (DG) and information only (IO) conditions were provided background reading materials adapted from Congressional Research Service and Congressional Budget Office reports. Below is a copy of the reading materials.

Please carefully read the following background information about immigration in the U.S. Afterwards you'll have the chance to provide your opinions on this topic.

INTRODUCTION

Non-citizens can enter the United States legally on a permanent basis, or on a temporary basis. If a person is granted permission to come into the country permanently, he or she is known as a legal immigrant and gets a "green card." In 2004, 362,000 people came into the United States this way. After five years, if they learn English and meet other conditions, legal immigrants can become citizens. About 537,000 people completed the process to earn citizenship in 2004. Non-citizens can also enter the country on a temporary visa, as a tourist, student, or temporary worker. These visitors are not expected to stay beyond the term of their visas. Anyone without a green card or a current visa is considered an illegal immigrant.

ILLEGAL IMMIGRANTS

About 12 million illegal immigrants live in the U.S., according to recent estimates. Every year, about half a million (500,000) new illegal immigrants enter the country. Between half and two-thirds come from Mexico. Sometimes crossing the border can be dangerous. Smugglers known as "coyotes" often use unsafe methods to sneak their customers across the border. The U.S Border Patrol believes that nearly two thousand people died trying to cross the border between 1998 and 2004.

California is home to the largest number of illegal immigrants, followed by Texas, Florida, New York, Arizona, Illinois, New Jersey, and North Carolina. Illegal immigrants can be deported if they are caught. In 2004, 1.2 million were caught; some left voluntarily while others were deported. Deporting illegal immigrants can be complicated if the immigrants have children who were born in the U.S., because under current law, these children are legal citizens, even if their parents are not.

ECONOMIC IMPLICATIONS

People are often concerned about how illegal immigration affects the job market, as well as taxes and social services like health care and education.

Right now, illegal immigrants make up about 5 percent of the U.S. work force. Many immigrants work in textiles, food manufacturing, construction, agriculture, food services, and janitorial services, where they earn 27 percent less than U.S. citizens with similar education and experience in the same industries. About 75 percent of the illegal immigrant population works. While it is very difficult to say with precision how illegal immigration affects wages, a report by the Congressional Budget Office suggests that it primarily affects American workers without high school diplomas. The wages for such jobs go down (by about 4 percent), which hurts these workers, but raises profits for American employers and businesses, and lowers prices for American consumers. Immigrants are consumers too, who pay for American products when they are here, so they contribute to the economy in that way as well. And some argue that immigration encourages American workers to invest in education to compete for higher-wage jobs.

TAXES & SOCIAL SERVICES

It is also difficult to know exactly how illegal immigration affects taxes and social services. Although many illegal workers pay social security and other taxes, they are not eligible for many government benefits. On the other hand, over a quarter of illegal immigrants live in poverty. Many use emergency health care, and their children attend U.S. schools (although some of those children were born here, and so are legal citizens who are entitled to public school education).

LEGISLATIVE EFFORTS

Taking on the issue of illegal immigration, both the House of Representatives and the Senate have passed legislation in recent months. The bills are very different, and in order to pass a law to set immigration policy, the two houses must come up with one bill that will pass in both chambers. Then the President must sign the bill to establish new immigration law.

The Senate bill, called the *Comprehensive Immigration Reform Act of 2006*, contains a path for illegal immigrants to become permanent residents if they pay a fine and go through a process to qualify as legal citizens. The bill also grants more visas to immigrants coming to work in certain industries where demand for their labor is higher (guest workers). Under current law, an American company who wants to use foreign workers under such programs must prove that doing so will not hurt the employment of current U.S. citizens.

The House of Representatives bill, called the *Border Protection, Antiterrorism, and Illegal Immigration Control Act of 2005*, makes it a felony to be in the United

States without proper documentation. Under this proposal, anyone who knowingly helps illegal immigrants can be prosecuted for a felony as well.

DETAILS OF THE SENATE BILL

Temporary Guest Worker Program

When certain industries have high demand for workers, this bill sets up temporary visas for workers to come to this country to get jobs in those industries. These guest workers must have a job lined up before they enter the country. They can stay up to three years, and can bring their families with them. No illegal immigrants currently living and working in the U.S. would be eligible for this program. For the first year, 325,000 workers could enter the country under this program. After that, the number would be adjusted every year, depending on the demand for workers in each industry.

Path to Citizenship

Illegal immigrants currently in the U.S. are not eligible for the guest worker program, but they may be allowed to become legal permanent residents. The bill sets up three different categories of illegal immigrants: those who have been in the country 5 years or more, those who have been here for 2-5 years, and those who have been here less than 2 years. The immigrants who have been here longest, since 2001 or earlier, can become permanent residents if they have been working for at least three of the five years. They have to pay a $5,000 fine. Their spouses and children will also get green cards. Once they have their green cards, they can eventually become citizens if they decide to go through that process too. Immigrants who came after 2001 and before 2004 (have been here 2-5 years) can get permission to stay and work for three years, provided they also pay a fine, of

$1,000, and have been working already for the last two years. With their new three-year visa, they can apply for other visas that allow for longer stays. To do this, they have to go to a point of entry on the border and file their application there.

Immigrants who have been in the U.S. less than two years will not receive any opportunities in the guest worker programs or paths to citizenship. They have to go back to their countries of origin and compete for a visa like everyone else.

Employer Sanctions

Under the Senate bill, fines for employers who knowingly hire illegal immigrants would be raised from their current amounts to $20,000. Repeat offenders would get jail time. Within 18 months, all employers would be required to use a database to verify that their employees are legal.

Border security

This bill would call for 370 miles of fencing along the U.S.-Mexico border, and another 500 miles of vehicle barriers. The Border Patrol, which has 11,000 agents right now, would be increased by 1,000 agents right away, and by 14,000 by 2011, for a total of 25,000 agents. The National Guard currently assists at the border, but under this bill, there would be a limit of 21 days to National Guard assignments there, to free up Guard troops when they are needed elsewhere.

English as the national language

The Senate bill establishes English as the official national language of the United States.

DETAILS OF THE HOUSE BILL

Border Security

The House Bill provides money for guarding the border with satellites, sensors in the ground, cameras, and radar. It also calls for 700 miles of fences along the U.S.-Mexico border, and more border patrol agents to patrol the fences.

Illegal entry and smuggling

Anyone caught smuggling illegal immigrants into the country can be prosecuted for aggravated felony charges, and could face mandatory minimum prison sentences. The bill also makes it a felony to be in the United States illegally. Immigrants face prison for entering the U.S. without proper documentation, and those who do so more than once face mandatory minimum prison sentences. People who marry illegal immigrants to help them get green cards face criminal penalties. So does anyone else who helps an illegal immigrant commit immigration fraud.

Employer sanctions

The House bill calls for fines of as much as $40,000 each time an employer hires an undocumented worker. Repeat offenders could face as much as 30 years of prison time. Within 6 years, employers would have to use a database to check Social Security numbers for each employee.

Sources: Congressional Research Service Reports for Congress: “Immigration: Policy Considerations Related to Guest Worker Program (October 2005);” “Immigration Legislation and Issues in the 109th Congress (January 2006);” Congressional Budget Office Papers: “Immigration Policy in the United States (February 2006);” “The Role of Immigrants in the U.S. Labor Market (November 2005);” Congressional Research Service Summary of bills H.R.4437 and S.2611.

B.3 KN and Deviations from an Ideal Design

As we note above, there were three main limitations to our experimental design that resulted from negotiations with KN: 1) administering an RSVP filter in the baseline survey; 2) excluding chronic nonresponders from the follow up and November surveys; and 3) subcontracting with other online polling firms for subjects. We assess each issue here in more detail and, for the first two, report sensitivity tests to assess the conceivable impact of the limitation on our reported results.

B.3.1 Retaining the RSVP "No" Respondents in the Analysis

As we describe above, we were certain that some noncompliance with the treatment would occur in our experiment, but neither we nor KN had data to estimate *ex ante* what the noncompliance rate would be. As a result, we faced a significant risk of assigning either too few or too many to the treatment condition. In order to ensure that we had sufficient numbers of subjects in each experimental cell (that is, for the treatment actually received), we included an RSVP filter question prior to randomization to gain information on the extent of noncompliance we would observe. Those who self-reported that they could or would not attend a session, but wanted to remain involved with the study, were randomized into one of the two control groups.

Provided that respondents who selected out of the treatment were truthful[4], then including this filter introduced no additional amount of noncompliance or adds any complexity to the statistical model that is designed to address problems of noncompliance; the same

[4] Note that the converse of this assumption need not be true, since the model assumes noncompliance. In addition, we check on the consequences of violations of this assumption below using sensitivity tests.

noncompliance would have occurred whether or not we introduced this filter. These self-reporting noncompliers would have ended up in a control group either way.[5] And as we describe in the paper, it is this noncompliance that the statistical method of principal stratification is designed to accommodate in identifying causal effects. Using this filter question to improve our assignment rates introduces no additional complexities to the analysis, but improves the power of the statistical model considerably since it provided very useful information on which to base the assignment rates.

The main limitation to this approach is that it gives us one fewer opportunity to observe the compliance behavior among those who otherwise would have been assigned to a deliberative session. In the statistical model that we report in the paper, we do not treat the self-report that one would not attend as behavioral data (i.e., as if this self-report were equivalent to actually not attending a session), since the self-report does not have the same status as a behavior. Using a model that is very similar to that of figure 1 (in the paper), we are able to retrieve the correlation between an RSVP "No" and the choice not to attend a session.[6] We estimate this correlation to be only 0.113 (with a 95 percent confidence interval of 0.078 to 0.143), which suggests that participants are declining the invitation for exogenous reasons, unrelated to their compliance type. As a consequence, in the paper we estimate subjects' compliance type based exclusively on their observed behavior and ignore their RSVP response.

[5] We note from Table A that the pretreatment covariate marginals between those who select out at the filter stage (and hence can only be in one of the control groups) and those who do not (but end up in one of the control groups) are nearly identical.

[6] We estimate this by re-running the model of figure 1 but this time including the RSVP response as another indicator variable for the compliance type. The correlations among the indicators are identified via the common latent variable (see Aakvik et al., 2005).

We are able to test the sensitivity of our estimated treatment effects to assumptions regarding the behavioral status of the RSVP response in our measure of the compliance type for these subjects. We also test for the sensitivity of our estimated effects to the range of conceivable responses to the knowledge items these subjects may have shown, had they been given a chance to attend a session.

We first test the sensitivity of the results we report in the paper to the assumption that the RSVP response does not indicate compliance type. As we mention above, the RSVP indicator does not load strongly on the compliance latent variable, and hence is virtually uncorrelated with the other compliance indicators. In a direct test, we re-estimated the main statistial model assuming that an RSVP "No" is the same as having been invited to a session and not showing up (i.e., treating the RSVP "No" as a behavioral response). The results of the model change only in the slightest, with the structural parameter for the deliberative group effect declining from 0.6 (standard error of 0.1) to 0.5 (standard error of 0.1). Likewise, the structural parameter for the information only condition declined from 0.3 (0.1) to 0.2 (0.1).

Next, we tested the sensitivity of the results to extreme assumptions about what knowledge gains the subjects that RSVP'd "No" might have had in response to exposure to a deliberative session, had they had access to the session and, contrary to their RSVP, actually attended. We re-estimated the model under the scenarios that some large number of those who RSVP'd "No" were assigned to the DG, actually attended a session and had a zero treatment effect. Specifically, we re-estimated the model under the scenario that we had 1) assigned 75 percent of those who RSVP'd "No" to the deliberative group, 2) that each and every such subject attended a session (no noncompliance, even though they had reported they would not attend), 3) and that each had a zero treatment effect (their post-treatment answers were identical to their pretreatment answers). Even under this unlikely scenario, we still would have found a significant and positive treatment effect for the DG treatment, with a structural parameter of 0.2 ($p < 0.05$). Of course, it would be odd to observe such a high rate of compliance among people who said they would not attend. In addition, Table A2 shows that those who responded "No" to the RSVP have very similar covariates to those who respond "Yes," so observing a zero treatment effect for these subjects would be surprising. We estimate this extreme scenario as a way to probe the outer limits of how our assumptions could conceivably affect our results. The sensitivity tests show that our results do not vanish even under very implausible assumptions of the RSVP "No" respondents behavior and treatment effects.

Finally, we re-ran the basic model (graphed in figure 1 of the paper) but where we drop out the 600 respondents who indicated they were unwilling or unable to attend the deliberative session. This restriction does not change the basic findings of our paper; the magnitude of the treatment effect is identical compared to the information only group and to the true control (and the differences remain statistically significant). The only difference is the treatment effect between the information only group and the true controls is no longer statistically significant, but this is to be expected as we have reduced the number of subjects in these two groups considerably and hence reduced the power of the comparison between them. This is further evidence that including the filter at the assignment stage has no effect on our main results beyond the noncompliance we observe independent of this filter. That the treatment effects within principal stratification remain the same, whether including or excluding this subgroup of noncompliers, strongly indicates that there is only one compliance mechanism (other than random noise), which the statistical model accommodates. That is, principal stratificiation only requires that those self-reporting they would not participate were being truthful and that, in fact, they would not have attended the session had they been invited.[7]

B.3.2 Excluding the Chronic Nonresponders from the Follow Up Survey

About halfway through the study, in negotiations with KN, we agreed to not send the follow up survey to those subjects who were least likely to respond: those in the information only condition and those in the deliberative condition who complied with none of their assigned tasks. Since this revision to the survey administration occurred halfway through the experiment, there were only 269 such subjects out of 2222, or about 12 percent of cases. Because we had already implemented the study in six districts (about half of the sample), 349 subjects who fall into this category had already been administered the follow up survey. Of these, only 25.7 percent actually responded to the follow up survey. Thus, assuming the order of the districts is unrelated to the compliance rate among this subset of subjects, had we sent the follow up survey to these remaining 269 chronic nonresponders, we would have expected to receive about 67 of them in return, or about 3 percent of the total sample. As we describe in appendix C.5, our statistical model is designed to impute the responses of these subjects, and does so on the basis of a rich set of data such as these subjects' pretreatment immigration policy knowledge, general political knowledge, and covariates such as education. We note that chronic nonresponders are likely less motivated to encode information on this topic, so if a bias exists in the imputation, the bias is likely to be against finding treatment effects.

We emphasize that the model does not assume these subjects would not have responded to the follow up survey had it been offered. Instead, our model makes use of all of their observed

[7] Note that the converse need not hold true, that those who initially reported they would attend need not be truthful. This is the ordinary noncompliance that principal stratification is designed to accommodate.

compliance and pretreatment covariates to impute the probability distribution of their response behavior, as well as the probability distribution for the responses they would have given on the post-treatment immigration policy knowledge outcome measures. In the model, note that pretreatment general political knowledge (η_4), pretreatment immigration policy knowledge (η_1), and compliance type (η_3) help to predict latent post treatment immigration policy knowledge (η_2), and that the missing responses for the knowledge items are imputed based on their mesaured post-treatment policy knowledge η_2. As we show in appendix C.4, η_2 by construction is a very strong predictor of responses to the knowledge items. That is, the statistical model makes use of a very rich information set to impute missing values for nonresponders. Since these imputations are in the form of distributions instead of as point estimates, our uncertainty over these imputations are propagated throughout the model, and hence appropriately increase the amount of uncertainty in the estimates of all structural parameters (Tanner and Wong, 1987).

Since the model makes this imputation based on observed compliance and pretreatment covariates, the imputation for the chronic nonresponders is no more and no less difficult than for others who choose not to respond on the outcome survey. This is reinforced empirically in Table A2 (below) where we show the covariates are closely balanced between those in the true control condition who chose not to respond and the chronic nonresponders, who end up as true controls and who were not given the chance to respond (see the final two columns of the table). An omnibus balance test (Hansen and Bowers, 2008) confirms the similarity of these two groups, as the joint distribution of the main covariates relevant to knowledge gains from deliberation (the count of the DC-K items correct, college or more, both need for cognition items, and both need for judgment items; see below) are equivalent (imbalance cannot be rejected, $p = 0.47$).

Just as we did with the RSVP filter, we checked the sensitivity of our estimated treatment effects to extreme assumptions of how the chronic nonresponders might have answered the knowledge items, had they been given the chance. Again, a sensitivity test is intended to see how robust our findings are, even if the chronic nonresponders had responded, and responded in very unexpected ways. In this case, we re-estimated the model, but under the assumptions that the chronic nonresponders 1) all responded to the follow up survey, and 2) had very large increases over their pretreatment knowledge, equal to that of the DG subjects, which is considerably higher than that of the controls who did return surveys. This would be both an unexpectedly high response rate and unexpectedly high knowledge gain for this subset of subjects, since nonresponders are least likely to be motivated to either return a survey or encode information related to the study. Under this scenario, our estimated treatment effect for the the deliberative group treatment is reduced only trivially, with the structural parameter changing from 0.6 (0.1) to 0.5 (0.1), and similar for the treatment effect for the information only condition, with the structural parameter changing from 0.3 (0.1) to 0.2 (0.1). This is likely because the relevant subset of respondents is such a small percent of the total sample. Thus, even under these extreme assumptions of how these chronic nonresponders would have responded, the treatment effects remain. We are very confident that the choice not to send this subset of subjects the follow up survey has no material consequence for our reported findings.

B.3.3 District Panel Sizes and Subcontracting for Subjects

As we mention in the text, Knowledge Networks maintains panels of potential survey respondents that are demographically representative. Because of the effort and resources required to maintain these panels, however, the panels themselves are relatively small. Since our study blocked on congressional districts (all treatment and control subject came from the

12 congressional districts in our study), KN's panels were not large enough in each congressional district to meet our size requirements. As a result, KN subcontracted with two other high quality online survey vendors, Survey Sampling International or SSI (http://www.surveysampling.com/en/methodologies/online-sampling) and Global Market Insite or GMI (http://www.gmi-mr.com/). While both of these subcontracting vendors maintain high quality panels, neither goes to the same lengths as KN to maintain representative samples. Since the panels that SSI and GMI maintain are very similar, we use panel fixed effects (KN versus SSI/GMI) to account for the differences between KN panels and the other two panels.[8]

That KN recruited some of our subjects from these other vendors means that the inferences we make in this study are limited to the population of subjects who join online survey panels, which themselves reflect the larger population that is well-connected to the Internet. We certainly cannot make inferences about how the average American would respond to our experiment if exposed to it, particularly those who stand on the other side of the digital divide. We do not see this as much of a limitation. If members of Congress were to one day adopt online town halls more broadly, it is the connected population who would likely be the ones to attend, and our sample is well-designed to assess the impacts on this population. And even if our study were limited to KN panelists, we likely would make this restriction in any case.

[8] Using separate fixed effects for SSI and GMI never yielded significant differences, so we collapse these to a single category.

C The Statistical Model

The statistical model we use in the paper is designed to address complications in data collection that one commonly encounters in conducting randomized field experiments (Gerber and Green, 2000). We rely on one major approach in the statistics literature to evaluate experimental data with noncompliance and nonresponse, the method of *principal stratification* coupled with a Bayesian parametric response model, developed by Donald Rubin and a number of his coauthors (Barnard et al., 2003; Frangakis and Rubin, 1999, 2002; Mealli et al., 2004). Our paper is a relatively straightforward application of this method, with some improvements to Rubin's methods that are enabled by our research design (as we describe in more detail below, we observe a variety of behaviors that help us pin down the key variable for principal stratification: the "compliance type" of the subject). Using alternative approaches, such as matching or instrumental variables, returns nearly identical point estimates, but standard errors that are considerably smaller. Hence, principal stratification is the most conservative approach for estimation.

C.1 Identifying Causal Effects using Principal Stratification

In the method of principal stratification (Frangakis and Rubin, 1999), the key to identifying treatment effects is to measure each subject's "compliance type," or the unobserved propensity to take up the treatment when offered, and hold this constant in the estimation. When the outcome model conditions on this latent compliance type variable, the outcome is made conditionally independent of both the treatment subjects actually receive and the pattern of missing outcomes, and the treatment effect estimate can be taken as causal.

The Frangakis and Rubin (FR) approach to principal stratification identifies treatment effects by assuming that the outcome is independent of both 1) treatment noncompliance and 2) the pattern of missing outcome data within strata of a categorical compliance variable. FR label this conditional independence assumption “latent ignorability.”[9] This conditioning variable classifies experimental subjects by their propensity to comply with the treatment. Under the assumption of latent ignorability (that treatment noncompliance, nonresponse on outcome measures, and the knowledge outcomes are independent within strata of a latent compliance variable), conditioning on the compliance variable compares likely compliers in the treatment group with likely compliers in the appropriate control group, and it compares likely noncompliers in the treatment group with likely noncompliers in the control group. In this way, principal stratification identifies causal effects (Barnard et al., 2003; Frangakis and Rubin, 1999; Horiuchi et al., 2007).

Principal stratification requires a measure of subjects’ latent compliance type. In the FR approach, however, principal stratification retains the significant limitation that it must treat compliance type as missing data for those assigned to the control group, since those assigned to the control typically do not have the opportunity to reveal their compliance type. This approach ignores any information on the compliance behavior of controls that is potentially available through the study. We implement principal stratification using a parametric item response model (e.g., Trier and Jackman, 2008) to measure the latent compliance type for all subjects, including those assigned to the control group (Esterling et al., 2011). In this

[9] Our implementation of the causal model requires four additional assumptions to identify causal effects: (1) randomization of a large number of subjects across the treatment and control groups ensures that the full range of compliance types resides in each treatment group; (2) monotonicity requires that the probability of compliance increases as the compliance type latent variable increases; (3) always takers do not exist, that those who

experiment, treatment as well as control subjects are asked to complete a series of assigned tasks. As each subject complies with or refuses each assigned task, she generates behavioral data that indicate her compliance type. Our implementation of principal stratification exploits these additional indicators of compliance type by measuring compliance type as a latent variable and including this latent characteristic as a control variable in a full structural equation model.

Building on the random effect approach in Aakvik et al. (2005), we extend principal stratification by estimating each subject's latent compliance type in a measurement model. In contrast to the FR approach to principal stratification, which assumes compliance is only observed among those assigned to the active treatment, our implementation measures compliance for all subjects, including the control subjects. This measurement is based on observed compliance behavior instead of a model-based imputation that relies heavily on the explanatory power of covariates. Since the full model incorporates a latent variable sub-model, we are able to demonstrate the validity of the behavioral measures we use to estimate compliance type; we discuss this validity test in appendix C.4.1.

In summary, principal stratification can accommodate the main complications with data collection that one encounters in large scale field experiments: noncompliance with the treatment and nonresponse on outcome surveys, making a correct comparison between treatment and control groups by holding constant each subject's compliance type. Our implementation of principal stratification imputes all missing data conditioned on latent variables that by construction are strong predictors of the missing data, such as compliance

are not offered the treatment do not have access to the treatment; and (4) the exclusion restriction, that the randomized assignment itself does not affect knowledge outcomes.

type for the compliance indicators, pretreatment policy knowledge for the post treatment policy knowledge indicators, and so on.

Other candidates for our statistical test include nearest neighbor matching, instrumental variables, and ordinary regression. In this application, these alternative estimators yielded point estimates that were nearly identical to those from principal stratification, but each returned t-statistics between seven and nine, strongly suggesting the standard errors were biased downward. Principal stratification requires weaker assumptions for this application than these more familiar estimators, returns much wider standard errors, and consequently is the most conservative. See appendix C.6.

C.2 Estimation

We implement the statistical model in the MCMC sampling software `WinBUGS` (Spiegelhalter et al., 1996). This sampler sequentially uses Bayes' Rule to update the parameter values until the posterior distribution converges to a stationary distribution; the resulting marginal stationary posterior distributions serve as the parameter estimates (see Jackman, 2000). One can characterize the point estimates and standard errors of all structural parameters, as well as functions of these parameters such as those we report in the manuscript, using the resulting posterior distribution. We approximate maximum likelihood (ML) estimates by assigning flat priors for all parameters.

Below we reproduce the `WinBUGS` code for the basic model, graphically shown in figure 1 of the paper. The code contains the "computer" variable names. To correspond the computer variable names to those found in the manuscript, note first that the indicator variables for each of the four latent variables (η_1 to η_4) are grouped together for each of the four measurement models. For example, the indicators for post treatment immigration policy knowledge, or η_2, are Q1post to Q6post. The exogenous variables (found in the equations for the conditional values of η_2 and η_3) are all self explanatory and are described either in the paper or in the control variable section below.

For identification, we must scale each latent variable to one observed or measured variable; the choice here is arbitrary. We scale each of the knowledge variables (η_2, η_1, and η_4) to the first indicator by assigning a value of one for the corresponding factor coefficient for the latent variable. For the compliance latent variable, η_3, after some trial and error, we found that the empirically most stable scaling variable was η_2, where we set the coefficient to 0.1.[10] Each indicator is dichotomous, and so are modeled with probit response functions. Since the response functions have no natural scale, the variance of the latent variables are not identified; we follow standard practice in item response models and set these variances to one.

```
### BEGIN WinBUGS MODEL

model{

for (i in 1:n){ ###

## post treatment immigration policy knowledge, eta2:
Q1post[i]~dbern(p1)
p1[i]<-phi(constant[1] + 1*eta2[i])
Q2post[i]~dbern(p2)
p2[i]<-phi(constant[2] + lambda.eta2[1]*eta2[i])
Q3post[i]~dbern(p3)
p3[i]<-phi(constant[3] + lambda.eta2[2]*eta2[i])
Q4post[i]~dbern(p4)
p4[i]<-phi(constant[4] + lambda.eta2[3]*eta2[i])
Q5post[i]~dbern(p5)
p5[i]<-phi(constant[5] + lambda.eta2[4]*eta2[i])
Q6post[i]~dbern(p6)
p6[i]<-phi(constant[6] + lambda.eta2[5]*eta2[i])
```

[10] We used this scale, 0.1, so that the remaining factor coefficients for η_3 were in a reasonable range as a way to improve convergence.

```
# Pretreatment immigration policy knowledge, eta1:
Q1pre[i]~dbern(p8)
p8[i]<-phi(constant[8] + 1*eta1[i])
Q2pre[i]~dbern(p9)
p9[i]<-phi(constant[9] + lambda.eta1[1]*eta1[i])
Q3pre[i]~dbern(p10)
p10[i]<-phi(constant[10] + lambda.eta1[2]*eta1[i])
Q4pre[i]~dbern(p11)
p11[i]<-phi(constant[11] + lambda.eta1[3]*eta1[i])
Q5pre[i]~dbern(p12)
p12[i]<-phi(constant[12] + lambda.eta1[4]*eta1[i])
Q6pre[i]~dbern(p13)
p13[i]<-phi(constant[13] + lambda.eta1[5]*eta1[i])
Q7pre[i]~dbern(p14)
p14[i]<-phi(constant[14] + lambda.eta1[6]*eta1[i])

## Compliance indicators for measuring compliance type, eta3:
## note that eta3 is scaled in the equation for eta2, below.
This
## scaling is empirically more stable than using one of the
indicator
## equations to scale.
participate.delib.session[i]~dbern(p20)
p20[i]<-phi(constant[15] + lambda.eta3[1]*eta3[i])
complete.bgm[i]~dbern(p21)
p21[i]<-phi(constant[16] + lambda.eta3[2]*eta3[i])
complete.followup[i]~dbern(p22)
p22[i]<-phi(constant[17] + lambda.eta3[3]*eta3[i])
complete.nov.survey[i]~dbern(p23)
p23[i]<-phi(constant[18] + lambda.eta3[4]*eta3[i])

## Genearal knowledge scale, eta4:
DCK1[i]~dbern(p24)
p24[i]<-phi(constant[19] + 1*eta4[i])
DCK2[i]~dbern(p25)
p25[i]<-phi(constant[20] + lambda.eta4[1]*eta4[i])
DCK3[i]~dbern(p26)
p26[i]<-phi(constant[21] + lambda.eta4[2]*eta4[i])
DCK4[i]~dbern(p27)
p27[i]<-phi(constant[22] + lambda.eta4[3]*eta4[i])
DCK5[i]~dbern(p28)
p28[i]<-phi(constant[23] + lambda.eta4[4]*eta4[i])

}
```

```
## Distribution for latent variables.
for (j in 1:n){

eta3[j]~dnorm(mu.eta3[j],1)I(-4,4)
mu.eta3[j]<- g.kn[1]*kn[j] + g.eta4[1]*eta4[j]
+ g.collegeormore[1]*collegeormore[j]
+ g.white[1]*white[j] + g.female[1]*female[j] +
g.work[1]*work[j]
+ g.expertsession[1]*expertsession[j] + g.ncI[1]*ncI[j] +
g.ncII[1]*ncII[j]
+ g.njI[1]*njI[j] + g.njII[1]*njII[j]

eta2[j]~dnorm(mu.eta2[j],1)I(-4,4)
mu.eta2[j]<- g.treatment*treatment[j] + g.infoonly*infoonly[j]
+ 0.1*eta3[j]
+ g.eta1*eta1[j] + g.tehetero*eta4[j]*treatment[j] +
g.kn[2]*kn[j]
+ g.eta4[2]*eta4[j] + g.collegeormore[2]*collegeormore[j]
+ g.white[2]*white[j] + g.female[2]*female[j] +
g.work[2]*work[j]
+ g.expertsession[2]*expertsession[j] + g.ncI[2]*ncI[j] +
g.ncII[2]*ncII[j]
+ g.njI[2]*njI[j] + g.njII[2]*njII[j]

eta1[j]~dnorm(0,1)I(-4,4)
eta4[j]~dnorm(0,1)I(-4,4)

}

## Priors for parameters

for (m in 1:6) {
lambda.eta1[m]~dunif(0,100)
lambda.eta2[m]~dunif(0,100)
}
for (m in 1:4) {
lambda.eta3[m]~dunif(0,100)
lambda.eta4[m]~dunif(0,100)
}
for (h in 1:23) {
constant[h]~dnorm(0.0,0.001)
}
g.treatment~dnorm(0.0,0.001)
g.infoonly~dnorm(0.0,0.001)
g.tehetero~dnorm(0.0,0.001)
g.eta1~dunif(0,100)
```

```
for (h in 1:2) {
g.kn[h]~dnorm(0.0,0.001)
g.eta4[h]~dnorm(0.0,0.001)
g.collegeormore[h]~dnorm(0.0,0.001)
g.white[h]~dnorm(0.0,0.001)
g.female[h]~dnorm(0.0,0.001)
g.work[h]~dnorm(0.0,0.001)
g.ncI[h]~dnorm(0.0,0.001)
g.ncII[h]~dnorm(0.0,0.001)
g.njI[h]~dnorm(0.0,0.001)
g.njII[h]~dnorm(0.0,0.001)
g.expertsession[h]~dnorm(0.0,0.001)
}

### end

}
```

We run this model until convergence, assessed using the Gelman and Rubin (1992) diagnostic. We then drew 25,000 values from the stable posterior distribution for each of three MCMC chains, retaining one in every 75, to create simulated marginal posterior distributions of the parameters with 1,002 draws.

Because we measure the compliance types of all subjects, we are able to retrieve average treatment effects for the sample (rather than complier average causal effects or other estimands found in the literature) by examining coefficients from the conditional response function of η_2, the latent variable that measures post-treatment immigration policy knowledge. In this equation, true controls are the baseline category. As a result, the coefficient `g.treatment` captures the difference between deliberators and true controls; `g.treatment` minus `g.infoonly` captures the difference between deliberators and the information only group; and `g.infoonly` captures the difference between the information only group and the true controls.[11] We retrieve the estimated effects of the treatment on the knowledge indicators, shown in figure 3 of the paper, using the ordinary procedure for differences in probabilities in a probit response function.

C.3 Additional Control Variables

The paper describes the control variable *College or more*. This section describes additional control variables that are included in the statistical models to assist in identifying causal effects. In addition to our identification strategy that relies on principal stratification, we include these variables to further safeguard there is no dependence between complying with the treatment, responding to the survey, and the knowledge outcomes in the analysis. These eight additional control variables are listed in appendix table A1.

Luskin (1990) and Nadeau and Niemi (1995) examine demographic determinants of political knowledge and, on the basis of their findings, we condition on race, gender, and employment. We also control for subjects' need for cognition (Cacioppo et al., 1984), which we anticipated would be a large determinant of subjects' propensity to join a deliberative sample, using two items measured on the baseline survey. We created the *Need for cognition I* variable by coding those who report having opinions about most things or about almost everything as one, and everyone else as zero. We created the *Need for cognition II* variable by coding those who report liking responsibility for handling situations that require a lot of thinking either a lot or somewhat as one, and everyone else as zero. We also control for subjects' need for judgment (Bizer et al., 2004) using two more items measured in the baseline

[11] In the model of causal mechanisms, described in appendix C.7, we also need to integrate out the compliance latent variable, η_1, from the response function for the mediating variable. We do this using numeric integration in `R`.

survey. We created the *Need for judgment I* variable as one if subjects reported that the statement "It is very important to me to hold strong opinions" as extremely or somewhat characteristic of them, and everyone else as zero. We created the *Need for judgment II* variable by coding as one those who report that the statement "I often prefer to remain neutral about complex issues" is somewhat or extremely uncharacteristic, and zero for those who believe this statement is characteristic of them. These variables measuring need for cognition and judgment may drive selection into and engagement with the deliberative sessions and hence are important control variables.

Finally, to get large enough samples for each district, our survey vendor, Knowledge Networks, subcontracted with two other online survey vendors. To account for differences between KN panelists and these other panels (which empirically are very similar to each other), we condition all outcomes on whether the subject is from a KN panel as a fixed effect.

The statistical models of the paper include all of these as control variables in the post-treatment immigration policy knowledge (η_2) model and in the compliance type (η_3) model. Overall, these variables have few substantively or statistically significant effects. These non-findings are likely due to both equations accounting for education and prior political and policy knowledge.

Table A1: Additional Control Variables Included in the Model

	Mean	S.D.	N
KN Panelist	0.357	0.479	2222
White	0.814	0.389	2222
Female	0.698	0.459	2222
Works Fulltime	0.405	0.491	2222
Need for Cognition I	0.722	0.448	2222
Need for Cognition II	0.632	0.484	2222
Need for Judgment I	0.773	0.419	2222
Need for Judgment II	0.595	0.491	2222

We have discussed how our statistical model can identify treatment effects. Even if one were to believe that principal stratification cannot address complications from noncompliance and nonresponse, our model conditions on these variables as an additional safeguard to control for pretreatment immigration policy knowledge, as well as important other predictors of post-treatment knowledge, such as general political knowledge (η_4), college education, need for cognition, need for judgment, as well as a host of other pretreatment covariates. In addition, as we discuss below, all missing response and outcome data are imputed using these pretreatment measures. Thus, even if principal stratification does not fully correct for a correlation between the treatment and the potential outcomes (and response behavior), netting out pretreatment immigration policy knowledge as well as the other pretreatment knowledge variables almost certainly should.

Table A2 (in this appendix) gives a breakdown of subjects' pretreatment covariate averages by the treatment they actually received and their response behavior. The table also further breaks these averages down according to responses on the initial filter question. We present this table as descriptive information for the sample only; it is not intended to indicate balance among the different groups across these covariates, nor does the statistical model require balance.[12] We do note, however, that the raw data do not begin with any marked departures from balance across these groups, when considering respondents and nonrespondents separately. There is a slight difference on dimensions for which one would expect to observe imbalance between deliberators and controls, especially in college degree, works full time, and need for cognition and judgment. There are very few differences between those in the information only group and the true controls. There are differences across the board, however, between responders and nonresponders. Of course, as we describe above, our statistical model is designed to accommodate differences between responders and nonresponders; these differences simply counsel the necessity of not conducting a "complete case" analysis that treats the nonresponse mechanism as random or conditionally random.

Table A2 helps to assure us that the two design features we included to improve the power of the statistical analysis pose no problem for identifying treatment effects within the framework of principal stratification. First, we introduced an RSVP filter question for subjects to self-report they would not or could not attend a session to help ensure we would have a sufficient number of subjects in each experimental cell. The marginals between those who reported they would not attend a session (and so were eligible only for one of the control groups) look very similar to the those who indicated they would attend a session with minor exceptions for works full time and KN panelist.[13] Second, in our design, those assigned to either the deliberative condition or to the information only condition who comply with none of their assignments effectively reassign themselves to the true control for the treatment they actually receive. In the study, a subset of these subjects were not sent the follow up survey. Table A2 shows that the true controls who were not administered the follow up survey and true controls who chose not to respond have very similar averages for their pretreatment covariates.

[12] As we note above, the treatment is ignorable since the model controls for compliance type principal strata. In addition, the model controls for many other pretreatment variables.

[13] That those who opt out at the filter stage look similar to those who do not suggests that many of these subjects drop out for exogenous reasons such as scheduling conflicts, making much of the noncompliance ignorable. Again, this takes some of the pressure off of the statistical assumptions in identifying treatment effects.

Table A2: Descriptive Statistics by Treatment Actually Received

	Delib. Group		Info. Only		True Control		
	R	NR	R	NR	R	NR	NS
Num. DC-K Items Correct							
Able/Willing	3.9	3.8	3.6	3.6	3.7	3.1	3.5
Not Able/Willing	–	–	3.8	3.7	3.8	3.3	3.1
Combined	3.9	3.8	3.7	3.6	3.8	3.2	3.5
White							
Able/Willing	0.84	0.83	0.84	0.82	0.85	0.76	0.74
Not Able/Willing	–	–	0.85	0.86	0.85	0.88	0.83
Combined	0.84	0.83	0.84	0.83	0.85	0.83	0.75
Female							
Able/Willing	0.66	0.68	0.69	0.70	0.70	0.77	0.73
Not Able/Willing	–	–	0.63	0.82	0.65	0.74	0.74
Combined	0.66	0.68	0.67	0.72	0.67	0.75	0.73
Works Fulltime							
Able/Willing	0.47	0.44	0.45	0.46	0.47	0.42	0.43
Not Able/Willing	–	–	0.33	0.32	0.27	0.26	0.33
Combined	0.47	0.44	0.42	0.43	0.35	0.33	0.41
College or More							
Able/Willing	0.51	0.47	0.43	0.34	0.41	0.33	0.39
Not Able/Willing	–	–	0.48	0.46	0.45	0.35	0.32
Combined	0.51	0.47	0.47	0.44	0.36	0.43	0.34
Need for Cognition							
Able/Willing	0.79	0.76	0.68	0.67	0.68	0.66	0.66
Not Able/Willing	–	–	0.67	0.63	0.61	0.60	0.57
Combined	0.79	0.76	0.68	0.66	0.64	0.63	0.65
Need for Judgment							
Able/Willing	0.77	0.77	0.68	0.70	0.66	0.61	0.69
Not Able/Willing	–	–	0.66	0.71	0.65	0.60	0.58
Combined	0.77	0.77	0.67	0.70	0.65	0.61	0.67
KN Panelist							
Able/Willing	0.57	0.14	0.54	0.22	0.36	0.09	0.12
Not Able/Willing	–	–	0.64	0.21	0.60	0.27	0.25
Combined	0.57	0.14	0.57	0.22	0.51	0.19	0.14
Cell Sizes							
Able/Willing	322	113	265	110	173	97	542
Not Able/Willing	0	0	113	28	268	115	76
Combined	332	113	378	138	441	212	618

Notes: R="Responder" (Responded on Follow Up Survey); NR="Nonresponder" (Did not respond on Follow Up Survey); NS="No Survey" (Not offered the Follow Up Survey). "Ability/Willing" are subjects who indicated they would be able and willing to do a deliberative session if invited; "Not Able/Willing" indicates the self-reporting noncompliers.

C.4 Measurement Model Results

The structural equation model diagrammed in figure 1 of the paper includes four latent variables, each of which is estimated with a measurement model. Each measurement model is an item response model with dichotomous indicator variables (see Patz and Junker, 1999; Trier and Jackman, 2008). Since we are estimating the outcome variable in a measurement model, the model accounts for errors in measuring the individual responses (see Achen, 1975; Imai and Yamamoto, 2008). The full structural equation model regresses some latent variables on other latent variables as well as on measured covariates. Appendix table A3 presents the factor coefficients, λ, for each measurement model.

In appendix table A3, cells give the estimated factor coefficients for each measurement model, along with standard errors. All coefficients are statistically significant at $p < 0.05$. Statistically significant coefficients show the reliability of each indicator. The compliance and the post-treatment immigration policy knowledge indicators are observed only among subsets of subjects; missing indicators are imputed in the model under the conditional independence assumption that is standard for latent variable models, as described in appendix C.5.

Table A3: Factor Coefficients for the Latent Variables

	λ	S.E.(λ)
Pretreatment Immigration Policy Knowledge (η_1)		
Question 1 Correct	1	0
Question 2 Correct	0.545	0.063
Question 3 Correct	0.144	0.044
Question 4 Correct	0.336	0.046
Question 5 Correct	0.159	0.042
Question 6 Correct	0.539	0.056
Post-Treatment Immigration Policy Knowledge (η_2)		
Question 1 Correct	1	0
Question 2 Correct	0.422	0.049
Question 3 Correct	0.195	0.038
Question 4 Correct	0.335	0.042
Question 5 Correct	0.131	0.036
Question 6 Correct	0.587	0.060
Compliance Indicators (η_3)		
η_2	1	0
Particip. in Delib. Gp.	0.542	0.052
Resp. BGM Survey	1.318	0.176
Resp. Followup Surv.	1.015	0.124
Resp. Nov. Survey	0.838	0.120
Political Knowledge (η_4)		
Cheney’s Current Job	1	0
Branch Determ. Const.	1.173	0.100
Maj. to Override Veto	0.660	0.051
Current Maj. Party	0.574	0.049
Party More Conserv.	1.201	0.106

C.4.1 Validity Test of the Compliance Indicators

The statistical model combines four separate measurement models, one for each latent variable. Using the factor coefficients for each measurement model, one can retrieve the correlation between indicators of each measurement model, and this enables a formal statistical test of the validity of the indicators (see Barnard et al., 2003, 305). Since the items are modeled using probit likelihoods, the formula to derive the inter-item correlations is,

$$\rho_{O_j,O_k} = \frac{\lambda_j \lambda_k}{\sqrt{\left(\lambda_j^2+1\right)\left(\lambda_k^2+1\right)}} \qquad (1)$$

where *j* and *k* indicate two distinct items. Using this formula, we find the correlation between participating in a discussion and responding to the BGM survey is 0.375 (95 percent confidence interval, 0.308 0.447); between participating in a discussion and responding to the follow up survey is 0.336 (95 percent confidence interval, 0.272 0.406); and between participating in a discussion and responding to the November survey is 0.301 (95 percent confidence interval, 0.238 0.374). That is, all of the indicators for responses to surveys are correlated with the main compliance indicator, participating in a deliberative session, and hence are valid for constructing the compliance scale.

C.5 Imputing Missing Data for the Endogenous Variables

It is widely recognized that discarding observations that contain missing data may cause biased estimates in any statistical method unless the data happen to be missing completely at random (MCAR) (e.g., Barnard et al., 2003). Frangakis and Rubin (1999) propose the method of principal stratification to address the problems of both noncompliance and missing outcome data. Under the latent ignorability assumption,[14] the missing knowledge outcome responses and the compliance indicators are conditionally independent within strata of the compliance type variable, and hence imputation through data augmentation enables unbiased estimates of treatment effects. This conditional independence assumption is standard in latent variable models, such as IRT models (see Trier and Jackman, 2008). With data augmentation, the

[14] As we discuss above, latent ignorability assumes that compliance type is correlated both with the outcomes and with the missing data process.

uncertainty inherent in the imputation is propagated through the posterior model parameters (Tanner and Wong, 1987).

C.6 Comparing Principal Stratification to Matching and Instrumental Variables

The statistical model that we implement in this paper is admittedly complex, particularly when compared to the methods one ordinarily encounters in the literature on causal effect estimation, such as nonparametric matching and instrumental variables (IV) estimation. Simplicity certainly has its virtues, but there are very good reasons why principal stratification is most appropriate for our estimation problem. In this application, principal stratification, matching, and IV estimate identical treatment effect estimates. For these data, however, principal stratification makes weaker assumptions for identifying causal effects than do the traditional methods and, as a consequence, it returns the most conservative estimates for the standard errors among the various estimators.

To show this, we created factor scores for subjects' post-treatment immigration policy knowledge (η_2), pretreatment immigration policy knowledge (η_1), and general political knowledge (η_4) using the same indicators we use to estimate these latent variables in the principal stratification model. For the matching study, we constructed an estimated propensity score by regressing each subject's decision to comply with the treatment (among those assigned to the treatment condition) on pretreatment immigration policy knowledge, general political knowledge, the covariates listed in the paper, and fixed effects for each congressional district in a logit equation. Using the nearest neighbor matching a do file available in Stata (Abadie et al., 2001), we estimate a sample average treatment effect of one half of a standard deviation increase in post-treatment immigration policy knowledge when comparing those in the deliberation group to true controls, and about a third of a standard deviation increase compared to those in the information only condition. These point estimates are nearly identical to those from principal stratification. But the matching z-scores are 6.5 and 4.2, respectively, indicating standard errors that are considerably smaller than those from principal stratification.[15]

We get very similar results using IV estimation. IV estimates the local average treatment effect (LATE), or the effect of the treatment on subjects who were induced to take up the treatment by their assignment (i.e., the treatment effect among the compliers), and this estimand is typically larger than the average treatment effect since it considers the effect among those on whom the experiment has the greatest impact. We estimate a LATE of a one full standard deviation increase when comparing the deliberators to the true controls, and slightly more than a half of a standard deviation increase when compared to the information only subjects. The z-scores for these estimate are 9.1 and 4.8, respectively, again showing that principal stratification is the more conservative estimator. We also conducted the IV analysis discarding the subjects who self-reported in the pretreatment survey that they would not attend a session even if invited, and the IV estimates are identical to those from the full sample. This latter finding offers still further evidence that there was no statistical consequence to allowing subjects to self-report their noncompliance (as opposed to revealing their noncompliance during the experiment).

[15] Other matching algorithms yield similar results; for example, the `GenMatch` algorithm (Diamond and Sekhon, 2007) returns a z-score of 4.5.

C.7 Identifying Direct and Indirect Effects in the “Causal Mechanism” Model

The model we outline in figure A2 demonstrates the effect of the treatment on a mediating variable (discussing immigration policy with others outside of the experiment), which in turn has an effect on the main outcome (immigration policy knowledge). In addition, this model tests whether the treatment has an effect on the outcome holding the mediating variable constant (i.e., does participating in the sessions themselves improve knowledge beyond what would be predicted by the values of the mediating variable?). In statistical parlance, the former is known as an “indirect” effect and the latter a “direct” effect (Mealli and Rubin, 2003; Rubin, 2004). That is, the model has three paths for treatment effects, testing 1) whether the treatment has an effect on the mediating variable, 2) whether the mediating variable has an effect on the final outcome, and 3) whether the treatment has an independent effect on the final outcome.

Specifically, we regress each of the dichotomous intervening variables on the indicator for deliberative group, the indicator for information only group, and the compliance latent variable, running the model separately for each of the two intervening variables. To test whether these indicators of cognitive motivation have a causal effect on policy knowledge, we simultaneously regress the latent immigration policy knowledge variable, η_2, on each indicator, holding constant compliance type, general political knowledge, pretreatment immigration policy knowledge, and the exogenous covariates used in the first model. Importantly, we also include two interaction variables, one that interacts the external attention indicator with the indicator for participating in the deliberative group, and one that interacts the external attention indicator with the indicator for the information only group. These interaction variables test whether attention to external information is disproportionately beneficial for one or another of the experimental groups. In particular, one might expect that those who participate in a deliberative session may feel more accountable for their immigration policy attitudes, to the extent that they anticipate discussing their experience in the session with others, and as a result attend to external information more closely, and hence encode the information more substantially.

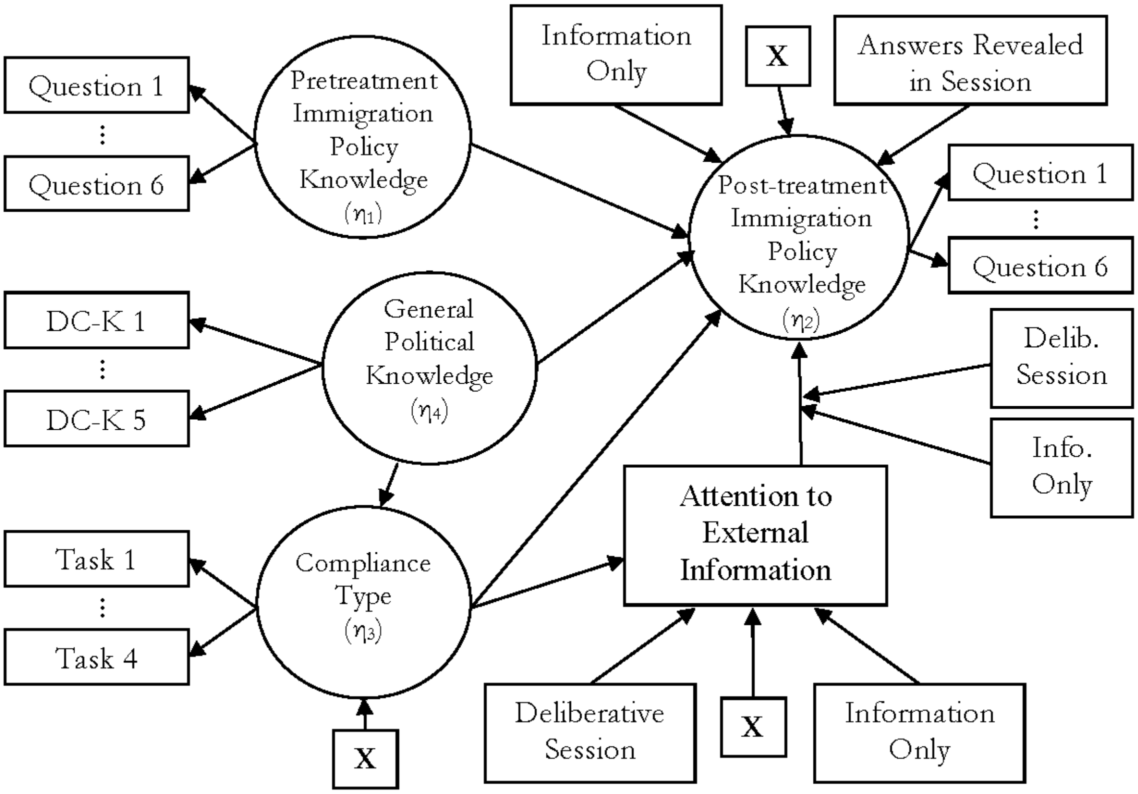
Question 1
Question 6
Pretreatment Immigration Policy Knowledge (η_1)
Information Only
X
Answers Revealed in Session
Post-treatment Immigration Policy Knowledge (η_2)
Question 1
Question 6
DC-K 1
DC-K 5
General Political Knowledge (η_4)
Delib. Session
Info. Only
Task 1
Task 4
Compliance Type (η_3)
Attention to External Information
X
Deliberative Session
X
Information Only

Figure A2: Casual Mechanisms

Notes: We estimate two versions of this model. In the first, the attention to external information variable is the subject's post-treatment belief that she has a duty to keep informed about politics. In the second, the attention to external information variable is the subject's self-report of whether she discussed immigration policy with others outside of the experiment.

A model testing a mediating effect necessarily relies on a stronger assumption, sequential ignorability, which requires conditional independence between the treatment and the mediating variable, after conditioning on covariates and the latent variables (Imai et al., 2010). The key for estimation is to model each of these paths within the framework of principal stratification. The model shown in figure A2 makes it apparent that the model controls for the compliance type in both the intermediate outcome equation (discusses immigration policy with others outside of the experiment) as well as in the final outcome equation (post-treatment immigration policy knowledge), hence the treatment, the mediating variable, the outcome variable, and the response pattern can be taken as conditionally independent within the assumptions of principal stratification (see the full discussion of principal stratification above).

C.8 Modeling the Compliance Process

The statistical model also conditions the compliance type latent variable η_3 on the same covariates that are included in the immigration policy knowledge equation, as well as general political knowledge. These results are interesting in themselves as they suggest the determinants of who chooses to deliberate in this kind of session, and we report a few of the results below. In this equation, we find that the coefficient relating general political knowledge to compliance type, γ_{43}, is positive and statistically significant, but substantively very small. A one standard deviation increase in general political knowledge only leads to a 0.1 standard deviation change in the compliance propensity (with 95 percent confidence interval of 0.0 to 0.2). The coefficient on the indicator for college education is not statistically significant. Thus, participation in a deliberative session, and interest in the study more generally, is only weakly related to subjects' prior political knowledge.

D Scaling up the Experimental Design

The results of the deliberative experiments involving members of the House of Representatives and their constituents are very encouraging. In brief, we find that citizens do seem to have the capacity and the willingness to gain knowledge when given a reason to do so, or in this case, when given the opportunity to engage in a direct act of accountability on an important issue. As we mention above, this capacity and willingness to gain knowledge is important to deliberative theorists in evaluating the health of a democracy. A skeptic might argue, however, that the positive effects that we find for our deliberative exercises are limited only to small group sessions. To address this concern, we briefly discuss a second experiment in which a large group of citizens from Michigan were given the opportunity to discuss enemy combatant detainee policy with Senator Carl Levin (D – MI) in July of 2008. We do not discuss this experiment in the same detail as the larger experiment involving members of the House of Representatives. Our goal here is simply to explore the issue of scalability in online deliberation experiments.

The Levin experiment had a similar design to the experiment with House members, but with a few differences. First, instead of small groups of 8 to 30 participants, the Levin session had 193 subjects who participated in the deliberative session. In addition, 327 subjects participated in the information only condition and 379 were true controls for a total of 899 subjects. Second, the topic for discussion was detainee policy, a far less salient issue than immigration policy, particularly at the times each study was conducted.[16] Third, we ran this experiment through the online survey firm Polimetrix.[17] Fourth, the deliberative session lasted 45 minutes, and there was no open-ended chat among participants after the session. Finally, due to funding constraints, we combined the baseline and background materials surveys, and all of the surveys were shorter.

The Levin study post-treatment (follow up) survey contained five items measuring detainee policy knowledge, listed in table A4. The percent correct rates for these items are very high. Recall that the immigration policy knowledge items were all nearly at the guess rate on the pretest. In contrast, the guess rates for detainee questions 1, 2, and 4 is 25 percent, and for items 3 and 5 is 33 percent.

The detainee knowledge items of table A4 do not have high inter-item correlations, so we estimate treatment effects for the individual items. In this experiment we combined several surveys and did not administer a post-election survey. As a consequence, we have too few compliance tasks to estimate a latent compliance variable (η_3 in the model above). Instead, we use the standard principal stratification Bayesian estimator (Frangakis and Rubin, 1999, 2002), implemented in Kosuke Imai's R package experiment.[18] For covariates, we include an indicator for whether the subject completed college (38 percent of the sample), an indicator

[16] Compare Dennis Jacobe, "Economy Widely Viewed as Most Important Problem," Gallup Poll, March 13, 2008, http://www.gallup.com/poll/104959/Economy-Widely-Viewed-Most-Important-Problem.aspx with Jeffrey M. Jones, "Immigration, Gas Prices Climb on Most Important Problem List," Gallup Poll, April 20, 2006, http://www.gallup.com/poll/22474/Immigration-Gas-Prices-Climb-Most-Important-Problem-List.aspx.

[17] See http:\\www.polimetrix.com.

[18] NoncompLI function, "Bayesian Analysis of Randomized Experiments with Noncompliance and Missing Outcomes Under the Assumption of Latent Ignorability," version 1.1-0. Documentation is available at http://imai.princeton.edu.

for whether the subject reports she is interested in news and public affairs “most of the time” (51 percent of the sample), and an indicator for whether the subject got all five of the Delli Carpini and Keeter items correct (46 percent of the sample).

We find significant treatment effects in this experiment, but on fewer items compared to the House study. Those who participated in the deliberative group were about 16 percentage points more likely to answer the first question correctly than those who were in the information only condition, regarding which conventions the U.S. has signed, with a 95 percent confidence interval of 0.0 to 33 percent. Those who participated in the deliberative group were also 16 percentage points more likely to answer this first item correctly compared to those in the true control group (with 95 percent confidence interval 4 percent to 27 percent), and about 14 percentage points more likely than true controls to answer the third item correctly, regarding the legality of torture (with 95 percent confidence interval of 6 percent to 22 percent). Overall, compared to the true controls, the point estimates for knowledge gains for the deliberative group were positive for all five of the detainee items, although only two of these effects were statistically significant given this sample size.

Table A4: Detainee Policy Knowledge Items

	Question	Response Set
1	Do you happen to know whether the U.S. has signed the Geneva Conventions (the international laws governing the treatment of people during wartime) and the United Nations Convention Against Torture?	a) Only the Geneva Conventions b) Only the United Nations Convention Against Torture c) **Both (44 percent correct)**
2	The U.S. military removed the Taliban from government in which country:	a) Iraq b) Saudi Arabia c) **Afghanistan (71 percent correct)** d) Israel
3	Under U.S law and the United Nations Convention Against Torture, torture is *legal*:	a) Only against citizens of countries who have *not* signed the treaty b) Only when both the President and a special court certify that there is a clear and present danger that requires it c) **Never, regardless of nationality, danger, or certification (79 percent correct)**
4	About how many captured people has the United States sent to the detention facility at Guantanamo Bay, Cuba?	a) 50 b) **500 (53 percent correct)** c) 2000 d) 5000
5	The Bush Administration has argued that the President has the authority to hold some individuals captured during combat against the U.S. without trial and indefinitely. Do you happen to know whether or how the Supreme Court ruled on this issue?	a) The Supreme Court has not ruled on the issue b) The Supreme Court has ruled in *support* of the Bush Administration's position c) **The Supreme Court has ruled *against* the Bush Administration's position (59 percent correct)**

Note: Boldface font indicates the correct answer and (pretreatment percent correct).

In this large group deliberative experiment, though knowledge gains are substantial we find fewer items with statistically significant treatment effects compare to the small group sessions involving House members. This may be for several reasons. It may be that large groups are less effective in inducing knowledge gains compared to small groups. We assert, however, that there are several more plausible explanations for the differences in treatment effects across the two experiments. First, the sample size of the study was smaller and hence the analysis simply had less power. Second, the topics in the two experiments were different, and we would not expect treatment effects to be constant across issues. Clearly detainee policy in the summer of 2008 was a far less salient issue than immigration policy in the summer of 2006. Citizens were perhaps more likely to be motivated to discuss immigration policy vigorously with others outside of the first experiment. Third, all subjects were more likely to give a correct response on the detainee items compared to the immigration policy items.[19] Whatever the reason for the different rates of correct answers, there was simply less room for improvement on the detainee items compared to the immigration policy items. On all three counts, one would have good reason to expect smaller treatment effects on the Levin experiment compared to the experiment involving House members.

The Levin experiment gives two important results. First, the online deliberative sessions are certainly scalable in a practical sense in that we had little difficulty in placing nearly 200 constituents in the online session and in managing the large group session. Second, the Levin session did produce knowledge gains even in a context, as noted above, where such gains might be especially difficult to induce.

E Conclusion

Since this study was a randomized experiment, one might be inclined to assess the design and execution of our field experiment according to the standards typically applied to laboratory

[19] Cost constraints for this study prevented us from using Polimetrix's match sampling procedures that would have generated a more representative sample similar to the Knowledge Networks sample.

experiments, where randomization is likely to be perfect. In such a setting, a statistical model need not do any "heavy lifting" at all. In large scale field experiments, however, one is virtually certain to encounter problems with noncompliance with the treatment and nonresponse on outcome measures. When using ordinary adults in experimental research, there simply is no ethical way to compel subjects to take up a treatment or respond on follow up surveys. The statistical model we use is designed to correct for departures from randomization instead at the analysis stage.

In return, however, field experiments can greatly improve the external validity of experimental findings. We feel very strongly that one can learn more about the effects of alternative institutions for deliberation effects and accountability processes using members of Congress and their constituents than using undergraduates in the lab, and confederates playing legislative roles, despite the inevitable complications that arise when undertaking a major experiment in the field. The statistical questions for data generated in a field experiment revolve less around whether the randomization of the treatment received is perfect, but instead whether statistical methods exist to identify causal effects with known departures from randomization. Just as the econometrics literature developed methods to overcome violations of OLS assumptions, there is an active literature in statistics that we rely on for how to identify causal treatment effects when randomization is imperfect.

References

Aakvik, Arlid, James J. Heckman and Edward J. Vytlacil. 2005. "Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics* 125(March):15–51.

Abadie, Alberto, David Drukker, Jane Leber Herr and Guido W. Imbens. 2001. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 1(1):1–18.

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69(Dec.):1218–1231.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98(462):299–323.

Bizer, George Y., Jon A. Krosnick, Allyson L. Holbrook, S. Christian Wheeler, Derek D. Rucker and Richard E. Petty. 2004. "The Impact of Personality on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate." *Journal of Personality* 72(Oct.):995–1027.

Cacioppo, John T., Richard E. Petty and Chuan Feng Kao. 1984. "The Efficient Assessment of Need for Cognition." *Journal of Personality Assessment* 48(May):306–307.

Callegaro, Mario and Charles Disogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72(5):1008–1032.

Diamond, Alexis and Jasjeet S. Sekhon. 2007. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." UC–Berkeley, Department of Political Science typescript.

Esterling, Kevin M, Michael A. Neblo and David M.J. Lazer. 2011. "Estimating Treatment Effects in the Presence of Selection on Unobservables: The Generalized Endogenous Treatment Model." forthcoming, *Political Analysis*.

Frangakis, Constantine E. and Donald B. Rubin. 1999. "Addressing Complications of Intention-toTreat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika* 86(2):365–379.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(March):21–29.

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7(Nov.):434–455.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(Sept.):653–663.

Hansen, Ben B. and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23(2):219–236.

Horiuchi, Yusaku, Kosuke Imai and Naoko Taniguchi. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51(July):669–687.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2010. Identification of Causal Mechanisms from Experimental and Observational Data. In *Third Annual West Coast Experiments Conference*.

Imai, Kosuke and Teppei Yamamoto. 2008. “Causal Inference with Measurement Error: Nonparametric Identification and Sensitivity Analysis of a Field Experiment on Democratic Deliberations.” Princeton University, Department of Politics typescript. http://imai.princeton.edu.

Jackman, Simon. 2000. “Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.” *American Journal of Political Science* 44(April):369–398.

Luskin, Robert C. 1990. “Explaining Political Sophistication.” *Political Behavior* 12(Dec.):331–361.

Mealli, Fabrizia, Guido W. Imbens, Salvatore Ferro and Annibale Biggeri. 2004. “Analyzing a Randomized Trial on Breast Self-Examination with Noncompliance and Missing Outcomes.” *Biostatistics* 5(2):207–222.

Mealli, Fabrizia and Donald B. Rubin. 2003. “Commentary: Assumptions Allowing the Estimation of Direct Causal Effects.” *Journal of Econometrics* 112:79–87.

Nadeau, Richard and Richard J. Niemi. 1995. “Educated Guesses: The Process of Answering Factual Questions in Surveys.” *Public Opinion Quarterly* 59(Autumn):323–346.

Patz, Richard J. and Brian W. Junker. 1999. “Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses.” *Journal of Educational and Behavioral Statistics* 24(Winter):342–366.

Rubin, Donald B. 2004. “Direct and Indirect Causal Effects via Potential Outcomes.” *Scandinavian Journal of Statistics* 31:161–170.

Spiegelhalter, David, Andrew Thomas, Nicky Best and Wally Gilks. 1996. BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual (version ii). Technical report MRC Biostatistics Unit.

Tanner, Martin A. and Wing Hung Wong. 1987. “The Calculation of Posterior Distributions by Data Augmentation.” *Journal of the American Statistical Association* 82(398):528–540.

Trier, Shawn and Simon Jackman. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science* 52(Jan.):201–217.