Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment

Yusaku Horiuchi   Australian National University
Kosuke Imai   Princeton University
Naoko Taniguchi   Teikyō University

Randomized experiments are becoming increasingly common in political science. Despite their well-known advantages over observational studies, randomized experiments are not free from complications. In particular, researchers often cannot force subjects to comply with treatment assignment and to provide the requested information. Furthermore, simple randomization of treatments remains the most commonly used method in the discipline even though more efficient procedures are available. Building on the recent statistical literature, we address these methodological issues by offering general recommendations for designing and analyzing randomized experiments to improve the validity and efficiency of causal inference. We also develop a new statistical methodology to explore causal heterogeneity. The proposed methods are applied to a survey experiment conducted during Japan's 2004 Upper House election, where randomly selected voters were encouraged to obtain policy information from political parties' websites. An R package is publicly available for implementing various methods useful for designing and analyzing randomized experiments.

I n this article, we demonstrate how to effectively design and analyze randomized experiments, which are becoming increasingly common in political science research (Druckman et al. 2006; McDermott 2002). Indeed, the number of articles in major political science journals that analyze randomized experiments has more than doubled since 1995.[1] Randomized experiments are more likely to yield unbiased estimates of causal effects than typical observational studies because the randomization of treatment makes the treatment and control groups equal on average in terms of *all* (observed and unobserved) characteristics.[2]

Yusaku Horiuchi is senior lecturer, Crawford School of Economics and Government, the ANU College of Asia and the Pacific, the Australian National University, Canberra, ACT 0200, Australia (yusaku.horiuchi@anu.edu.au). Kosuke Imai is assistant professor of politics, Princeton University, Princeton, NJ 08544-1012 (kimai@Princeton.Edu). Naoko Taniguchi is assistant professor of sociology, Teikyō University, 359 Ōtsuka, Hachiōji, Tōkyō 192-0395, Japan (n-oni@yb3.so-net.ne.jp).

[1]We have examined all the articles that were published since 1995 in *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*. From 1995 to 2000, only 14 articles using randomized experiments were published in these journals. This number increased to 35 during the next six years (from 2001 to 2006).

[2]Randomization also guarantees that the treatment is causally prior to the outcome thereby avoiding both posttreatment and simultaneity biases.

Even in randomized experiments, however, complications can arise especially when they are conducted outside of the laboratory (Barnard et al. 2003; Imai 2005). In political science experiments, researchers often cannot force subjects to comply with treatment assignment (noncompliance) and to provide the requested information (nonresponse). Since experimental subjects can *choose* not to comply and/or not to respond, ignoring this selection process leads to invalid causal inferences. Unfortunately, a majority of existing studies in political science ignore at least one of the two problems.

Furthermore, experimental designs used by political scientists remain elementary. Although many statistical textbooks on experimental designs recommend efficient randomization procedures such as randomized-block and matched-pair designs (e.g., Cochran and Cox 1957; Cox and Reid 2000), simple randomization of treatments remains the single most commonly used method in the discipline.[3]

In this article, we address these and other methodological issues. Building on the recent statistical literature, we show how to make statistical adjustments for noncompliance and nonresponse at the same time and illustrate how to design randomized experiments and conduct efficient randomization so that the potential impact of such complications is minimized. We also introduce a new statistical method and estimation technique to model causal heterogeneity in randomized experiments. Using such a method, researchers can learn more about underlying causal processes.

In addition to new statistical methods we propose, this article attempts to pave the way for further development of more methodologically sophisticated experimental studies in political science by offering four general recommendations. First, *researchers should obtain information about background characteristics of experimental subjects that can be used to predict their noncompliance, nonresponse, and the outcome*.[4] Although randomization is a powerful tool, it is incorrect to assume that there is no need to collect additional pretreatment variables. We show that without rich pretreatment information, researchers will not be able to obtain valid causal estimates when noncompliance and nonresponse problems exist.

The pretreatment covariates are also essential for efficient randomization and investigation of causal heterogeneity.

Second, *researchers should conduct efficient randomization of treatments by using, for example, randomized-block and matched-pair designs*. Randomized-block designs are implemented and discussed in this article. In matched-pair designs, experimental subjects are paired based on their values of the pretreatment covariates before randomization of the treatment. These randomization schemes are simple to implement but can yield more efficient estimation of causal effects than would be possible under the simple randomization design.

Third, *researchers must make every effort to record the precise treatment received by each experimental subject*. This information allows for greater accuracy in the interpretation of estimated causal effects as well as sensitivity analyses about modeling assumptions. In addition, we show that additional information about the administration of treatments can help researchers learn more about underlying causal processes through the direct modeling of causal heterogeneity.

Finally, *a valid statistical analysis of randomized experiments must properly account for noncompliance and nonresponse problems simultaneously*. Using analyses of empirical and simulation data, we show that the Bayesian framework of Imbens and Rubin (1997) provides a more flexible way to model such complications than the standard frequentist instrumental variable approach. Within this framework, we first build our base model which is closely related to the model proposed in the literature. We then demonstrate the flexibility of this framework by developing new methods to model causal heterogeneity in randomized experiments. C-code, along with an easy-to-use R interface (R Development Core Team 2006), is publicly available as an R package, *experiment* (Imai 2007), through the Comprehensive R Archive Network (http://cran.r-project.org/) for implementing our proposed and other methods that are useful for designing and analyzing randomized experiments.

Roadmap of the Article. Although this is an article about statistical methods, for the purpose of illustration, we have designed and conducted an Internet-based survey experiment during Japan's 2004 Upper House election. The second section briefly reviews relevant literature on information and voter turnout and provides some background information about Japanese politics. Readers who are only interested in methodological issues may skip this section and directly go to the third section where the design of our randomized experiment is presented and relevant methodological issues are discussed. In the fourth section, we review a statistical framework for analyzing randomized experiments with noncompliance and

---

[3]We find that almost all of the experimental studies identified in footnote 1 and listed at the Time-Sharing Experiments for the Social Sciences website (TESS; http://www.experimentcentral.org/) do not use these designs. The search via the Social Science Citation Index also yields no recent experimental study in political science that uses these techniques.

[4]This can be achieved, for example, by conducting a screening survey as we do in our experiment (see the third section).

nonresponse and show how to substantively interpret its underlying assumptions in the context of our illustrative example. The fifth section introduces our base model and further develops a new methodology to model causal heterogeneity. The results of our analysis are presented in the sixth section, and the seventh section offers concluding remarks.

An Experimental Study of Information and Voter Turnout

Since the publication of Downs's influential book in 1957, political scientists have emphasized information as a key determinant of voting behavior (e.g., Alvarez 1998; Ferejohn and Kuklinski 1990; Grofman 1993). In this article, we focus on Downs's argument about the effects of information on voter turnout. Downs started with the assumption that "Many citizens . . . are uncertain about how to vote. Either they have not made up their minds at all, or they have reached some decision but feel that further information might alter it" (1957, 85). Downs further argued that these voters are likely to abstain if they remain uncertain on election day, but that they are most susceptible to persuasion through additional information (85–86). The above argument suggests two simple hypotheses: (1) the more information voters receive, the less likely they are to abstain, and (2) additional information increases turnout especially for those who are uncertain about which party/candidate to vote for.[5]

To empirically test these hypotheses, we designed a randomized experiment during Japan's 2004 Upper House election where randomly selected voters were encouraged to view the designated official website of the Liberal Democratic Party (LDP) and/or the Democratic Party of Japan (DPJ).[6] Specifically, we used a particular section of their official policy proposals, so-called manifestos, about pension reform. The pension reform was seen as the major issue in this election[7] and fits Downs's argument that voters are likely to pay attention to information about "contested policies," "new policies," and "new situations" (1957, 217). Indeed, one month before the election, a new pension legislation was passed in the Diet. This legislation was widely considered to be a set of minor changes to the pension system, and opposition parties contested various proposals for additional reform during the campaign period. The pension reform was also necessary to meet new situations in Japan—rapidly aging population and rising budget deficit.[8]

Japan is an ideal case to test the Downsian hypotheses because a large number of voters are uncertain about how to vote; in Downs's terms, they are either "baffleds" or "quasi-informed neutrals" rather than "loyalists" who always vote for the same party or "apathetics" who always abstain. Many researchers consider Japanese voters to have relatively weak partisanship (e.g., Richardson 1997). Moreover, recent research highlights the growing number of Japanese voters with no party identification, known as *mutōhasō*. The percentage of these voters has increased from approximately 30% in 1988 to about 50% in the late 1990s (Matsumoto 2006; Weisberg and Tanaka 2001). Since then, the proportion of *mutōhasō* has been consistently larger than that of supporters for any political party. Researchers also have shown that these voters are not necessarily apathetic or apolitical. Rather, they are interested in politics, but consciously choose not to support any party or choose which party to support at each election (Matsumoto 2006; Tanaka 1997).[9] These studies imply, according to the Downsian hypothesis, that Japanese voters are likely to be susceptible to additional information (see also Basinger and Lavine 2005).

---

[5]While we empirically examine two Downsian hypotheses, we do not directly test the causal mechanism of information effects and hence do not answer the questions of why information increases turnout and how much information is necessary to alter turnout. Also, we neither compare across different kinds of information nor address the issue of how information effects may vary across different electoral institutions. These questions are important but beyond the scope of this article.

[6]The LDP has been in power since its foundation in 1955 (except between 1993 and 1994), and the DPJ is the largest opposition party, formed by amalgamation of various parties. In the 2004 Upper House election, in which we designed our experiment, over 80% of seats were won by these two major parties.

[7]According to a poll (*Asahi Shimbun*, evening edition, June 24, 2004) one week prior to the 2004 Upper House election, 53% of candidates and 47% of voters regarded pension reform as "the most important issue."

[8]The LDP web page explained in detail the new legislation, but gave little information about plans for further reforms, such as the widely debated future integration of various pension systems. It also did not mention other controversial issues, including the question of whether to abolish the special pension scheme available for Diet members. In contrast, at the very beginning of its party manifesto, the DPJ emphasized the need for national integration of pension systems and proposed the abolishment of the special pension. While the DPJ website proposed major reforms, it neither specifed the content of the reforms nor explained how such proposals would be implemented.

[9]According to the poll conducted one week prior to the 2004 Upper House election (see footnote 7), over 50% of self-identified *mutōhasō* were undecided about which party/candidate to vote for, whereas less than 20% of respondents who support a particular party were undecided.

Furthermore, Japan also provides an interesting case because major Japanese political parties have recently begun to prepare manifestos that explicitly state their formal policy proposals. This change was brought about as a part of the recent political reform to promote policy-based electoral competition. Political parties use these manifestos to attract more voters by offering additional information about their policy proposals to voters during the campaign period. In addition, Japanese parties and candidates are not allowed to change the contents of their websites during the campaign period. This regulation provides a methodological advantage for our experimental study, as we can guarantee that the contents of our treatment remain the same during our study.

Finally, while randomized experiments have their own limitations, there is a significant advantage when estimating causal effects of information on voting behavior. Namely, an experimental study allows us to manipulate the quantity and quality of additional information each voter receives. Survey researchers have designed various questions to measure the amount of information voters possess (e.g., Bartels 1996). Although the literature yielded considerable insight about the possible *association* between voting and information, it faces a common methodological challenge that hinders the estimation of the *causal effect* of information. The problem is that those voters who have a strong intention to vote may be more likely to acquire the information (e.g., Prior 2005). Randomized experiments can address this problem of *endogenous information acquisition* by giving additional or different information to randomly selected voters.[10]

Designing Randomized Experiments

In this section, we describe the design of the randomized experiment we conducted through a Japanese Internet survey firm, *Nikkei Research*, in June and July of 2004.[11] The purpose of this section is to illustrate, with this particular experiment, how to incorporate three out of our four methodological recommendations when designing randomized experiments: (1) to collect background

characteristics that are key predictors of noncompliance, nonresponse, and the outcome; (2) to conduct efficient and accurate randomization of treatments; and (3) to record the treatment received as precisely as possible. Later sections discuss in detail the remaining recommendation and its connection with the ones discussed here. Although experiments with different goals often require different designs, we believe that these methodological points are fairly general and can be incorporated into many experimental studies.

Collecting the Pretreatment Information

Our experiment consisted of three separate surveys: screening, preelection, and postelection surveys. Between June 24 and 29, two weeks prior to the election day (July 11), we conducted the screening survey and asked 6,000 randomly sampled respondents to answer several questions about themselves and their voting intention in the upcoming election.[12] The main purpose of this survey was to collect the background characteristics of experimental subjects, which are important predictors of noncompliance, nonresponse, and the outcome. Therefore, the first part of the survey questionnaire asked for each respondent's prefecture of residence, age, gender, and highest education completed. We also asked the respondents about their party preference and whether they were planning to vote in the upcoming election (planning to vote, not planning to vote, or undecided) and, if so, which party and/or candidate they were going to vote for, and how much confidence they had in their voting plan (a 4-point scale). We measured these variables because they are expected to be powerful predictors of the outcome variable, i.e., voter turnout. To avoid further statistical assumptions at the data analysis stage, the survey was designed so that respondents must answer all of the questions in order to complete the survey.

Of 6,000 individuals who received an electronic mail asking them to fill out the screening survey, 2,748 individuals completed the survey. From this group, we then randomly selected 2,000 eligible voters as our experimental sample. With a few exceptions of overrepresented urban prefectures, a comparison of the 2000 Census data and our sample shows no clear evidence of geographical sampling bias. Yet, as is the case for many other experimental studies, our sample is not representative of the Japanese

---

[10]Although conducting experiments is a powerful approach, a careful analysis of observational data with appropriate assumptions can also shed light on causal relationships (e.g., Lassen 2005; Sekhhon 2004).

[11]The company has a sampling pool of roughly 40,000 Internet users throughout Japan who have agreed to receive occasional electronic mail asking them to participate in online surveys. Those who fill out a survey questionnaire have a chance to win a gift certificate in the amount of approximately five to ten dollars.

[12]We asked *Nikkei Research* to select randomly an equal number of male and female Internet users, all of whom were between ages 20 and 59. The wording of our survey questions closely follows the standard protocol from large-scale Japanese national surveys such as the Japan Election Study projects.

electorate. The individuals in our sample were those who had Internet access, voluntarily registered themselves for the survey firm, and agreed to fill out the screening survey. The lack of representativeness and the unavailability of sampling weights are weaknesses of the Japanese Internet survey company used in our experiment. Some Internet survey firms in other countries may provide researchers with a more representative sample and/or sampling weights (e.g., Knowledge Networks in the United States). In the fifth section, we develop a new method that can potentially incorporate such additional information to address the problem of nonrepresentativeness.

Randomizing and Administering the Treatments

From July 5 to 9, we conducted the preelection survey to administer the randomized treatments. The election was held on July 11. Before sending an electronic mail soliciting their participation in the survey, we randomly assigned the treatments to the voters of our experimental sample. In particular, we considered two types of treatments and randomly divided the sample into three groups. The voters in the *one-party treatment group* were asked to visit the designated website of either the Liberal Democratic Party (LDP) or the Democratic Party of Japan (DPJ), while those in the *two-party treatment group* were asked to visit the websites of both parties. As explained earlier, for both LDP and DPJ, we used the official websites that show their policy proposals on the pension reform. Since parties and candidates are not allowed to change the web contents during a campaign period, the contents of treatment cannot vary across voters within each treatment group. No voter in the control group was asked to participate in the preelection survey. In general, a control group is essential for causal inference that requires counterfactual analysis.[13]

In order to randomly divide the sample into treatment and control groups, we applied a randomized-block design shown in Table 1. We formed six blocks on the basis of the gender and voting intention variables, which we obtained from the screening survey. These variables are selected because they are important predictors of Japanese voters' turnout decision. Within each randomized block, we conducted the *complete* randomization (rather than the simple randomization) of treatments such that the total number of voters is 1,000, 600, and 400 for the

one-party and two-party treatment groups, and the control group, respectively. Within the one-party treatment group, we randomly selected half of the voters and instructed them to visit the DPJ website. The other half was instructed to visit the LDP website. For the two-party treatment group, a random half of the voters was instructed to visit the DPJ website first before visiting the LDP website, while the order was reversed for the other half.

An important advantage of the randomized-block design is that it effectively reduces random *and* systematic differences between the treatment and control groups along the coordinates defined by selected pretreatment covariates. Imai, King, and Stuart (2007) proves that the resulting estimates under the randomized-block design are more efficient than those obtained without any blocking (see also Cochran and Cox 1957). We recommend that researchers select the covariates that are good predictors of the outcome variable when forming randomized blocks. Other types of experimental designs based on similar ideas are also possible. For example, a matched-pair design, which is a special case of randomized-block designs, creates pairs of observations with similar characteristics (e.g., through a matching method based on the Mahalanobis distance measure) and conducts complete randomization within such pairs (e.g., Greevy et al. 2004; Hill, Rubin, and Thomas 1999). In general, these designs cannot decrease statistical efficiency (see Imai, King, and Stuart 2007; Greevy et al. 2004, Section 1 and references therein).

Before being instructed to visit the website(s), voters were presented with "warm-up" questions about the pension reform. After answering them, voters were instructed to click a direct link to the designated party website. The instruction also included a friendly warning, which mentioned they would be asked about their opinion about the website after visiting it. We designed the survey so that voters would have to visit the website before answering the next question. In addition, we obtained information about whether and how long voters actually opened the designated website in their browser even when the voters decided not to go to the next question.

Knowing the treatment received by each subject as precisely as possible is an important consideration because the interpretation of estimated causal effects is difficult without such information. In our experiment, we are interested in estimating the causal effect of being exposed to policy information on the official party website(s). However, if one wishes to know why visiting the party website(s) increases/decreases turnout, it is equally important to check whether the administered treatments actually correspond to the concepts that

[13]When the problem of noncompliance exists, a direct comparison of different treatment groups is difficult because compliers and noncompliers must be identified separately for each treatment (Cheng and Small 2006; Imai 2005).

TABLE 1   Randomized-Block Design of the Japanese Election Experiment

	Randomized Blocks						
	I	II	III	IV	V	VI	
	Planning to Vote		Not Planning to Vote		Undecided		
	Male	Female	Male	Female	Male	Female	Total
One-party treatment group							
DPJ website	194	151	24	33	36	62	500
LDP website	194	151	24	33	36	62	500
Two-party treatment group							
DPJ/LDP websites	117	91	15	20	20	37	300
LDP/DPJ websites	117	91	15	20	20	37	300
Control group							
no website	156	121	19	26	29	49	400
Block size	778	605	97	132	141	247	2000

Notes: Six randomized blocks were formed on the basis of the two pretreatment covariates, gender (male or female), and the answer to the question, "Are you going to vote in the upcoming election?" Within each block, the complete random assignment of the treatments is conducted so that the size of each treatment and control group is equal to the predetermined number. The total sample size is 2,000.

underlie theoretical models. This can be done by asking respondents additional survey questions. For example, in order to know whether respondents actually acquired policy information from party website(s), one could ask knowledge test questions about party policies. While the decomposition of causal effects into multiple causal pathways is an important topic for future research (e.g., Pearl 2000), its comprehensive treatment is beyond the scope of our article. Nevertheless, we partially address this question by modeling causal heterogeneity in the fifth section.

Finally, to complete the survey, voters were asked to answer several brief questions about the website they had just visited. For those voters who were assigned to the two-party websites, the same set of questions was presented after they visited each website. At the end of the survey, voters were given a chance to write freely their opinions about the website(s). Although this was optional, nearly 80% of those who participated in the preelection survey wrote some comments, indicating a strong interest in the pension reform, the upcoming election, and/or the party websites.

Measuring the Outcome Variable

The postelection survey was started on July 12, the day after the election, and was closed on the 16th. The goal was to measure turnout for all 2,000 experimental subjects. We used the same questionnaire for everyone, asking whether they had voted in the election. We kept the survey short

to minimize the unit nonresponse, yielding over 80% of the response rate.

Analyzing Randomized Experiments

In political science experiments, researchers often do not have full control over their human subjects, and as a result, noncompliance and nonresponse often coexist. In our experiment, some voters did not visit the designated website(s) even when they were instructed to do so (noncompliance). Moreover, some voters did not fill out the postelection survey, and therefore the outcome variable was not recorded (nonresponse). Since these two problems do not occur completely at random, ignoring either or both of them in estimation may severely bias causal inference.

If data are not missing completely at random, the simple mean-difference of the observed outcome between the treatment and control groups no longer produces a valid estimate of the causal effect of the treatment assignment. This is true even when the treatment assignment is randomized *and* when the missing data mechanism is not affected by the assignment and receipt of treatment, because list-wise deletion will necessarily change the population for which the causal effects are being estimated. Moreover, the instrumental variable method, which relies on the consistent estimation of the causal effect of treatment assignment, needs to be modified in the presence of noncompliance (Frangakis and Rubin 1999; Imai 2006).

In this section, we review the general statistical framework of randomized experiments with noncompliance and nonresponse, which was first introduced by Angrist, Imbens, and Rubin (1996) and was later generalized by Frangakis and Rubin (2002). Within this framework, we use our Japanese election experiment as a running example and illustrate how to substantively interpret causal quantities of interest and understand necessary statistical assumptions for analysis of randomized experiments with noncompliance and nonresponse.

Noncompliance

We begin by describing the formal statistical framework of Angrist, Imbens, and Rubin (1996) in the context of our Japanese election experiment. Formally, let Z_i be the treatment assignment variable, which equals 1 if voter i is instructed to visit a party website(s) and is 0 otherwise. Use $T_i(z)$ to represent the *potential* treatment receipt variable given the treatment assignment $Z_i = z$. The *actual* treatment receipt, which equals 1 if voter i actually visits the website(s) and is 0 otherwise, can then be defined as $T_i \equiv Z_i T_i(1) + (1 - Z_i) T_i(0)$. If a voter logs on to the preelection survey questionnaire website but does not visit the designated website(s), we set $Z_i = 1$ and $T_i(1) = 0$, and thus $T_i = 0$ for the voter.[14] There were 133 such individuals, 63 of which belong to the one-party treatment group. This corresponds to about 10% of the voters who logged on to the survey website.[15]

We now define potential outcome variables given the assignment of the treatment; $Y_i(z) \equiv Y_i(Z_i = z)$ which is equal to 1 if voter i turned out and 0 otherwise. The actual outcome is then defined as $Y_i \equiv Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.[16] For example, in the treatment assignment group, $Y_i(1)$ is observable but $Y_i(0)$ is not. We also define two types of individuals in our experiment—compliers and noncompliers. Compliers refers to the voters who visit the designated party website(s) *only when* instructed to

do so ($T_i(1) = 1$ and $T_i(0) = 0$), while noncompliers are those who do not follow the instructions. There are three types of noncompliers; *always-takers*, who visit the party website regardless of whether they are instructed to do so ($T_i(1) = T_i(0) = 1$), *never-takers*, who do not visit the party website regardless of the instruction ($T_i(1) = T_i(0) = 0$), and *defiers*, who visit the website only when they are not instructed to do so ($T_i(1) = 0$ and $T_i(0) = 1$). We use C_i as the complier indicator variable, which equals 1 if respondent i is a complier and is 0 if he is a noncomplier.

Angrist, Imbens, and Rubin (1996) assumes that *there is no defier* (monotonicity). In addition, we assume no always-taker. Since we were unable to obtain whether the individuals in the control group accessed these specific party websites, these assumptions are not testable from the observed data. The assumptions, however, appear to be reasonable in our experiment because only a small number of voters would have looked at the party website spontaneously. The survey conducted by *Asahi Shimbun* and the University of Tokyo shows that only 3% of its respondents visited websites of parties and candidates during the 2003 Lower House election.[17] It is also unlikely for ordinary voters to obtain an official manifesto. Japan's Public Office Election Law (Kōshoku Senkyo Hō) allows only limited distribution of manifestos by political parties. Indeed, another survey conducted by Fuji Sōgo Kenkyūjyo shows that even during the 2003 Lower House election, which was the first manifesto election in Japan and attracted a great deal of media attention, only 6% of respondents obtained a complete version of a manifesto while 10% saw its outline.[18]

Figure 1 shows that from the observed data (T_i, Z_i), we can identify the compliance status of the voters in the treatment assignment group (those in the upper and lower left cells of the figure). However, for the voters in the control group (those in the lower-right cell of the figure), we need to infer their compliance status using the observed compliance pattern of the treatment assignment group. As we shall see, randomization of treatment assignment Z_i makes such inference possible because it guarantees that voters in the control group are similar to those in the treatment assignment group in terms of their observed and unobserved characteristics.

Finally, we assume *no direct effect of treatment assignment* (exclusion restriction). For voter i, the treatment assignment Z_i is assumed to affect the voter's potential

[14] Because of such drop-outs, we conduct sensitivity analyses using three different definitions of treatment receipt (logged on to the survey website, visited the party website(s), and completed the survey). The detailed information about treatment receipt we collect allows for such comprehensive sensitivity analyses.

[15] In the two-party treatment group, there were 18 voters who visited only one designated website and did not complete the preelection survey. We set $T_i = 0$ for these voters.

[16] Implicit in this formulation of potential outcomes is the assumption of *no interference among units* (Cox 1958). In our case, the assumption is reasonable because the respondents are unlikely to communicate with each other.

[17] See http://politics.j.u-tokyo.ac.jp/data/data01.html (accessed on 12 June 2005).

[18] See http://www.mizuho-ir.co.jp/research/documents/manifesto 031113_report.pdf (accessed on 12 June 2005).

FIGURE 1   Compliers and Noncompliers in the Japanese Election
Experiment

| | Treatment Assignment | |
	$Z_i = 1$	$Z_i = 0$
Actual Treatment Received $T_i = 1$	**Complier** $Y_i(1)$ is observable; $C_i = 1$	
$T_i = 0$	**Noncomplier** $Y_i(1)$ is observable; $C_i = 0$	**Complier** or **Noncomplier** $Y_i(0)$ is observable; $C_i =$??

Notes: The figure classifies compliers and noncompliers by the actual treatment received, T_i, and treatment assignment, Z_i. We assume that noncompliers solely consist of never-takers. From the observed data (T_i, Z_i), one can identify compliers and noncompliers in all but one case where $T_i = Z_i = 0$. The upper right cell is empty because we assume that always-takers and defiers do not exist in this experiment.

outcomes $Y_i(Z_i)$ only through the actual treatment received T_i. This means that the causal effect is zero for never-takers because their treatment receipt is the same (i.e., $T_i(1) = T_i(0) = 0$) regardless of their treatment assignment. (The assumption can also be made for always-takers if they exist.) In our experiment, this assumption is violated if a respondent changes her decision to vote because she is instructed to visit the party website(s) even though she does not actually complete the preelection survey. This scenario is a potential concern for 113 voters who logged on to the preelection survey but did not complete it. Therefore, we conduct sensitivity analyses by applying various definitions of actual treatment received. This point directly relates to the third of our four methodological recommendations that we set forth earlier.

In our experimental sample, the proportion of compliers is estimated to be about 70% (see the sixth section), which is much higher than typical field experiments in political science (Imai 2005). A high compliance rate is crucial for successful statistical analyses of randomized experiments because it reduces the degree to which the validity of estimated causal effects relies on the aforementioned and other assumptions that are often difficult to verify from observed data.

Nonresponse

In many randomized experiments, the problem of non-response coexists with the noncompliance problem. In our Japanese election experiment, the nonresponse problem arises because some respondents do not fill out the

postelection survey. In particular, 342 respondents out of 2,000 individuals did not fill out the postelection survey (75 of them belong to the control group; 113 of them are members of the one-party treatment group; and the remaining 154 voters belong to the two-party treatment group), and hence the value of the outcome variable is missing for approximately 17% of the experimental subjects.

Unfortunately, even in randomized experiments, deleting the observations with nonresponse, as is often done in political science research, results in both bias and inefficiency (Frangakis and Rubin 1999; Imai 2006). The list-wise deletion yields bias for two reasons. First, it can destroy the randomization (i.e., the treatment assignment group is no longer similar to the control group after list-wise deletion) unless the missing data mechanism is not influenced by the treatment assignment itself. Second, unless the data are *missing completely at random*, the list-wise deletion changes the target population for which causal effects are being estimated. This is true even when the missing data mechanism is not affected by the assignment and receipt of treatments because the respondents are in general systematically different from nonrespondents. In addition to bias, inefficiency arises from the information loss due to the exclusion of some observations from the analysis.

In the statistical literature, several approaches have been developed to deal with nonresponse and noncompliance at the same time (Imai 2006). First, we consider the missing at random (or MAR) approach (Little and Rubin 1987). Since nonresponse can be affected by the treatment assignment, we define the potential response indicator variable, $R_i(z)$ for $z \in \{0, 1\}$. For example,

$R_i(1) = 1$ and $R_i(0) = 0$ would mean that respondent i would answer the postelection survey if he is assigned to the treatment, but would not answer if he is in the control group. The observed indicator variable is given by $R_i \equiv Z_i R(1) + (1 - Z_i) R(0)$. Then, the MAR assumption can be formally stated as

$$\Pr(R_i(z) = 1 \mid Y_i(z) = 1, T_i(z) = t, Z_i = z, X_i = x)$$
$$= \Pr(R_i(z) = 1 \mid Y_i(z) = 0, T_i(z) = t, Z_i = z, X_i = x), \quad (1)$$

for $t, z \in \{0, 1\}$ where X_i represents a vector of the pretreatment covariates. The assumption implies that after conditioning on the (observed) treatment and encouragement variables as well as pretreatment covariates, the pattern of nonresponse is no longer systematically related to the outcome variable.

An alternative assumption is called the latent ignorability (LI) approach (Barnard et al. 2003; Frangakis and Rubin 1999), which is formally defined as

$$\Pr(R_i(z) = 1 \mid Y_i(z) = 1, C_i = c, Z_i = z, X_i = x)$$
$$= \Pr(R_i(z) = 1 \mid Y_i(z) = 0, C_i = c, Z_i = z, X_i = x), \quad (2)$$

for $c, z \in \{0, 1\}$. This assumption is slightly more general than the MAR because it conditions on the unobserved compliance covariate rather than the observed treatment variable.[19]

The third assumption is the nonignorability (NI) approach (Imai 2006). Unlike the MAR and LI assumptions, the NI approach is appropriate if the missing-data mechanism *directly* depends on the (possibly unobserved) values of the outcome variable itself. Formally,

$$\Pr(R_i(1) = 1 \mid T_i(1) = t, Y_i(1) = y, Z_i = 1, X_i = x)$$
$$= \Pr(R_i(0) = 1 \mid T_i(0) = t, Y_i(0) = y, Z_i = 0, X_i = x), \quad (3)$$

for $t, y \in \{0, 1\}$. The assumption implies that the missing data mechanism does not depend on the randomized encouragement once we condition on the outcome variable as well as the actual treatment received and observed pretreatment covariates.

The appropriateness of each assumption depends on the context of one's experiment, and sensitivity analysis plays an important role when examining the robustness of the resulting conclusions (Imai 2006). Nevertheless, whichever assumption is made about the missing-data mechanism, a key point is to include important predictors of the outcome variable in X_i so that these assumptions hold. Such variables also lead to efficient estimation of

causal effects because they help to accurately predict the missing values of compliance status and outcome variable. This underlies the first of our four methodological recommendations made in the first section. In our case, along with the background characteristics that are known to be associated with the voting behavior of the Japanese electorate, X_i includes voting intention variables measured two weeks prior to the election. Therefore, the rich covariate information makes the assumption plausible.

Finally, it is important to emphasize that our screening survey is designed to ensure that there is no missing value among the pretreatment variables. This design allows us to focus on the nonresponse problem in the outcome variable. In general, researchers may encounter missing values for both the pretreatment and outcome variables. Barnard et al. (2003) shows how to deal with randomized experiments with a complex pattern of missing data.

Causal Quantities of Interest

The first causal quantity of interest we consider is the (sample) intention-to-treat (ITT) effect,[20]

$$\mathrm{ITT} = \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)]. \quad (4)$$

In our experiment, the ITT effect is the causal effect of being asked to visit the party website rather than the effect of actually visiting the website. Although the ITT effect is not the effect of the actual treatment received, from the perspective of policymakers who want to increase voter turnout by using the Internet, the ITT effect may be of great interest because political parties can not force every voter to visit their website. Moreover, so long as the treatment assignment is properly randomized, the estimation of the ITT effect does not require the assumptions of *no always-taker* and *no direct effect of treatment assignment* because it does not require the identification of unobserved compliance status. Unfortunately, as discussed above, in the presence of nonresponse, the estimation of the ITT effect is no longer straightforward and requires additional assumptions about the missing-data mechanism. In fact, the simple difference-in-means estimator will no longer be unbiased (Frangakis and Rubin 1999).

From a social scientific perspective, one may be interested in estimating the effect of the actual treatment received T_i rather than the ITT effect. Given the

---

[19]The knowledge of (C_i, Z_i) implies that of (T_i, Z_i) whereas the latter does not imply the former (see Figure 1).

[20]The population ITT effect is given by $E[Y_i(1) - Y_i(0)]$.

assumptions about noncompliance and nonresponse, we can identify this causal effect (Angrist, Imbens, and Rubin 1996). By definition, noncompliers never receive the treatment and so the effect of the actual treatment received cannot be inferred from the observed data for this subgroup. Hence, we focus on the causal effect of the treatment for compliers who are in the upper left and lower right cells of Figure 1. Then, a key estimand is the (sample) complier average causal effect or CACE, which is defined as[21]

$$\text{CACE} = \frac{\sum_{i=1}^{N} C_i [Y_i(1) - Y_i(0)]}{\sum_{i=1}^{N} C_i}. \tag{5}$$

In our experiment, CACE defines the causal effect of information on voting behavior *for those who would visit the website only when told to do so*. It is important to note that CACE does not equal the usual sample average treatment effect, which is the average effect for the entire sample.

Another interpretation of CACE is that it represents the ITT effect for compliers. Since compliers always follow the treatment assignment, the causal effect of the actual treatment received is the same as that of treatment assignment for this subgroup. In contrast, the ITT effect for noncompliers is zero under our assumption that the treatment assignment does not directly affect the outcome variable. Therefore, CACE will be always larger than the ITT effect. The difference is large when the compliance rate is low. Earlier, we argued that always-takers are unlikely to exist in our experiment. Even when always-takers exist and are ignored in the analysis, the direction of this potential bias is known under our setting. While the estimation of ITT effect is unaffected, CACE will be *underestimated* because compliers may include always-takers whose causal effect is likely to be less than that of compliers. The size of this bias depends on the proportion of always-takers in the sample.

Modeling, Estimation, and Inference

Based on the statistical framework reviewed in the previous section, we develop models to analyze our Japanese election data while accounting for both noncompliance and nonresponse. Conducting a valid statistical analysis that makes proper adjustments for these problems is our fourth recommendation. We adopt the Bayesian modeling framework of Imbens and Rubin (1997) because, as

demonstrated by Barnard et al. (2003) and others, this framework allows one to deal with noncompliance, nonresponse, and other complexities in a flexible manner. We start by introducing our base model, which is based on the one in the literature (Hirano et al. 2000). We then contribute to this literature by extending our base model in order to directly model causal heterogeneity in randomized experiments when noncompliance and nonresponse are present. We also develop an efficient Markov chain Monte Carlo algorithm based on the idea of marginal augmentation to estimate the proposed models.

Base Model

We now describe our base model. First, we model the conditional probability of being a complier given each voter's observed covariates. We use the following binary probit model with linear predictors,

$$\Pr(C_i = 1 \mid X_i, \xi) = \Phi(X_i^\top \xi), \tag{6}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, X_i includes indicator variables for each of the randomized blocks, and ξ is the vector of coefficients.

Second, we model voter turnout given the compliance status C_i, the treatment assignment Z_i, and the observed covariates X_i (the actual treatment received is redundant given C_i and Z_i). Again, we use a binary probit model,

$$\Pr(Y_i = 1 \mid C_i, Z_i, X_i, \alpha, \beta, \gamma)$$
$$= \Phi[\alpha C_i Z_i + \beta C_i (1 - Z_i) + X_i^\top \gamma], \tag{7}$$

where α and β are the intercepts specific to compliers with and without the treatment, respectively, and γ is a vector of coefficients for X_i, which includes intercepts specific to each randomized block. The base category is noncompliers. Depending on the type of the outcome variable, researchers can choose an appropriate model (e.g., a Normal linear model for a continuous outcome variable).

The two probit models in equations (6) and (7) are combined to form the following complete-data likelihood function,

$$\prod_{i=1}^{N} \left[\Phi(X_i^\top \xi) \Phi(\alpha Z_i + \beta(1 - Z_i) + X_i^\top \gamma)^{Y_i} \right.$$
$$\left. \{1 - \Phi(\alpha Z_i + \beta(1 - Z_i) + X_i^\top \gamma)\}^{1-Y_i} \right]^{C_i}$$
$$\times \left[\{1 - \Phi(X_i^\top \xi)\} \Phi(X_i^\top \gamma)^{Y_i} \{1 - \Phi(X_i^\top \gamma)\}^{1-Y_i} \right]^{1-C_i}. \tag{8}$$

However, we cannot directly evaluate this likelihood function for two reasons. First, the compliance status C_i is not

---

[21]The population CACE is given by $E[Y_i(1) - Y_i(0) \mid C_i = 1]$.

observed for the voters in the control group. Second, the value of the outcome variable Y_i is missing for some observations. In situations such as this, it is natural to consider the method of data augmentation where missing data are "imputed" based on statistical models. In our Japanese election application, we consider the MAR assumption of equation 1 about the nonresponse mechanism.[22] Under this assumption, each iteration of our estimation procedure imputes the unobserved values of C_i and Y_i using the probit models with estimated coefficients, the pretreatment variables, and the treatment assignment. Based on the imputed values, we update the estimated coefficients. We repeat these steps until a satisfactory degree of convergence is achieved.

To conduct the Bayesian analysis, we assign two independent conjugate prior distributions on (α, β, γ) and ξ, both of which are multivariate normal distributions with a mean of zero and a variance of 100. To estimate the model, we sample from the joint posterior distribution via a Gibbs sampling algorithm. We use the marginal data augmentation algorithm of Imai and van Dyk (2005) to exploit the latent variable structure and speed up the convergence of the standard sampler. The details of the algorithm appear in the appendix. The Gibbs sampling algorithm produces Monte Carlo samples from the joint posterior distribution of the unobserved compliance status, the missing values of the outcome variable, and the model parameters. Given these posterior draws, the estimates of CACE and the ITT effect can be easily calculated, as can their uncertainty estimates. Our convergence diagnostics are based on multiple independent Markov chains initiated at overdispersed starting values (Gelman and Rubin 1992). The sampling algorithm is implemented by our own C code.

Modeling Causal Heterogeneity

The development of statistical methods for causal inference with experimental data is an active area of research. In this section, we extend our base model by developing new models of causal heterogeneity in randomized experiments with noncompliance and nonresponse. Causal heterogeneity is one of the most important and outstanding methodological issues because answers to many scientific questions require the estimation of varying treatment effects across observations (e.g., Angrist 2004; Hotz, Imbens, and Mortimer 2005).

Two Types of Causal Heterogeneity. In general, causal heterogeneity can arise in two ways. First, treatment effects can be heterogeneous if, given the identical treatment, some experimental units experience larger treatment effects than others. We call this type of heterogeneity *treatment effect heterogeneity*. For example, in our experiment, we may hypothesize that the magnitude of treatment effects depends on the characteristics of respondents (see the second section).

Second, causal heterogeneity can also arise if the treatment is administered differently for each experimental subject. For example, subjects may receive different levels or kinds of the treatment (see also Imai and van Dyk 2004). This second type is called *treatment heterogeneity* because it is induced by the differences in treatments themselves rather than differences in the characteristics of units. Although experimenters should minimize treatment heterogeneity when designing experiments, eliminating it is difficult in practice. For example, in many experiments, treatments are administered at different times and locations and/or by different people. Experimental subjects also may *choose* to receive the assigned treatment in different ways. In our experiment, respondents decide how much time they spend viewing the party website(s), the information which we have recorded. Such information about treatment heterogeneity may be difficult to obtain, but it allows researchers to better understand underlying causal processes.

The two sources of causal heterogeneity described above share a common statistical estimation problem. Namely, we wish to model the treatment effect as an unknown function of an observed variable, denoted by D_i, which we call the heterogeneity variable. The key difference between the two is the nature of this heterogeneity variable. In the case of treatment effect heterogeneity, on the one hand, D_i represents an observed characteristic of units. In the case of treatment heterogeneity, on the other hand, D_i measures the observed levels/kinds (or "doses") of the treatment. While the former is a pretreatment covariate observed *before* the randomization of treatment assignment, the latter describes treatments themselves and may be influenced by the treatment assignment.

After introducing a simple modeling approach that is only applicable to treatment effect heterogeneity, we propose a more general approach that can deal with both types of heterogeneity. One potential advantage of our general approach is that sampling weights, when available, can be used as the heterogeneity variable, D_i. By modeling

[22]The modeling strategies under the LI assumption (equation 2) and the NI assumption (equation 3) are described in detail in Barnard et al. (2003) and Imai (2006), respectively. These models are also implemented via an R package, *experiment* (Imai 2007).

treatment effect heterogeneity as a function of sampling weights, our method allows researchers to address the lack of representativeness in the experimental sample and improve the external validity of their conclusions.

A Simple Approach. The first approach we illustrate estimates causal effects separately for different subgroups using pretreatment variables in order to model treatment effect heterogeneity. This is easy to implement because it only requires researchers to fit the same model independently in subsets of the data. There are, however, a few drawbacks. First, the method becomes inefficient as the number of subgroups increases and the size of each subgroup becomes small. Second, this approach is not applicable when modeling treatment heterogeneity because only pretreatment variables can be used to define subgroups. This is because, if one uses an observed variable affected by random assignment of treatment to define subgroups, the treatment assignment is no longer random within the resulting subsets of the data.

A General Approach. A more general approach is to directly model treatment or treatment effect heterogeneity. In particular, we generalize our outcome model in equation (7),

$$\Pr(Y_i = 1 \mid C_i, Z_i, X_i, \alpha(D_i), \beta, \gamma)$$
$$= \Phi(\alpha(D_i) C_i Z_i + \beta C_i (1 - Z_i) + X_i^\top \gamma), \quad (9)$$

where $\alpha(D_i)$ is an unknown function of D_i to be estimated. Note that the compliance model stays identical to equation (6), and the complete-data likelihood function in equation (8) for this generalized model is also the same except that α is replaced by $\alpha(D_i)$.

We emphasize that D_i can be either a pretreatment variable or a variable measuring different levels/kinds of the treatment. In the latter case, however, an additional assumption is required in order to give a causal interpretation because D_i is measured after the randomization of the treatment assignment, Z_i. The assumption is that the heterogeneity variable is conditionally independent of the potential outcome variables after conditioning on the observed covariates for those units who received the treatment (D_i is measured only for them).[23] The assumption is automatically satisfied if, for example, the treatment levels are also randomized. In the absence of such randomization, however, the assumption implies that researchers must collect relevant pretreatment covariates in order to make the assumption plausible, especially when experimental subjects can choose different levels or kinds of

treatment.[24] This point is an important consideration, as emphasized in our first methodological recommendation in the first section.

When D_i is a discrete variable with a small number of distinct categories, equation (9) describes the basic model with additional interaction terms of the heterogeneity variable, D_i, and the treatment variable, T_i. Such a model only requires to add interaction terms to equation (7) and thus is straightforward to estimate in our framework. Here, we consider in detail a more general situation where D_i is a continuous variable or a discrete variable with a large number of ordered categories. One way to estimate the model in such situations is to assume that $\alpha(D_i)$ has a specific functional form, e.g., a linear function of D_i. However, such a functional form assumption may be too restrictive if one is interested in estimating the functional relationship between the treatment and the outcome from the data rather than assuming it.

Thus, we propose a method that only requires a very weak assumption using a semiparametric "differencing estimator" (e.g., Yatchew 1998). We do not specify a particular functional form of $\alpha(D_i)$. Instead, our approach only requires $\alpha(D_i)$ to be a *smooth* function. Formally,

$$|\alpha(D_{i'}) - \alpha(D_{i'+1})| \ < k \, | \, D_{i'} - D_{i'+1}|$$
$$\text{for} \quad i' = 1, \ldots, N_t - 1, \quad (10)$$

where k is an unknown small constant, and $N_t = \sum_{i=1}^{N} T_i$ treated units are ordered by the value of the heterogeneity variable, D_i, and are indexed by i' (such that $D_{i'} \leq D_{i'+1}$ for all $i' = 1, \ldots, N_t - 1$). The assumption implies that if the values of the heterogeneity variable are similar between any two units, then the effects of the treatment on those units should also be similar.

Estimation. This generalized model can be estimated by slightly modifying the Bayesian estimation technique introduced earlier. To do this, note that a vector of smooth coefficient $(\alpha(D_1), \ldots, \alpha(D_{N_t-1}))^\top$ can be reexpressed as $M \Delta$ where M is the known first-order differencing matrix, and $\Delta = (\alpha(D_2) - \alpha(D_1), \alpha(D_3) - \alpha(D_2), \ldots, \alpha(D_{N_t}) - \alpha(D_{N_t-1}))^\top$. Given this, the Bayesian model can be formulated by directly placing a conjugate prior distribution on the differences, Δ, rather than the effects themselves (Koop and Poirier 2004);

[23] When always-takers exist, D_i is measured for always-takers as well as compliers.

[24] Under alternative assumptions such as the constant additive treatment effect assumption, which may not be reasonable in many contexts, including ours, a different estimator can estimate the dose-response function (e.g., Angrist and Imbens 1995).

i.e., $\Delta \mid \sigma^2 \sim N(0, \sigma^2 I)$ and $\sigma^2 \sim \text{Inv } \chi^2(\nu_0, s_0^2)$, where ν_0 denotes the prior degrees of freedom parameter and s_0^2 represents the prior scale parameter. The degree of smoothness is controlled by σ^2 since a smaller value of σ^2 indicates a smoother function of $\alpha(D_i)$. For (β, γ) and the parameters of the compliance model, ξ, we use the same prior distributions as before. Then, the MCMC algorithm described in the appendix can be used to fit this generalized model.

Empirical Results

In this section, we present the results of our statistical analysis. Our inference is based on Monte Carlo samples from three independent Markov chains, each of which has the length of 50,000 and is initiated at different sets of starting values. We find that all the parameters have the values of the Gelman-Rubin convergence statistic that are less than 1.01, which suggests that a satisfactory degree of convergence has been achieved. We retain the last 10,000 draws from each chain and base our inference on a combined total of 30,000 posterior draws.

Results from the Base Model

First, Table 2 compares the estimated causal effects of one-party treatment with those of two-party treatment by presenting the posterior summaries of the ITT effect, CACE, and proportion of compliers in the sample. Also, the estimated turnout rates for the control group are shown separately for compliers, noncompliers, and the entire sample. They serve as the baseline turnout of no exposure to the designated party website(s).[25] Note that compliers are defined separately for each treatment; i.e., compliers for the one-party treatment are not identical to those for the two-party treatment. The estimated baseline turnout rate is about 70%, which is almost 15 percentage points higher than the official turnout rate of the 2004 Upper House election. This gap may arise because our sample is not representative of the Japanese electorate (see the third section).[26]

As expected from the hypothesis described earlier, the posterior means of CACE and the ITT effect are all positive. The estimated causal effects of the one-party treatment are small, and there is relatively large uncertainty. The estimated probability of voting increases (from 71%) by only about one percentage point on average among those voters who *would have visited a website if encouraged* (CACE), and the estimated increase is on average one percentage point (from 70%) among those who were *asked to visit* one website (the ITT effect).[27] Since the estimated proportion of compliers is quite high (75% in the one-party treatment and 70% in the two-party treatment), the estimated ITT effects and CACE are somewhat similar in our study. Large posterior standard deviations mean that one cannot statistically distinguish these small positive estimates from zero.

The estimated causal effects of the two-party treatment are larger, though the differences themselves are not statistically significant. The estimated turnout probability increases (from 70%) by three percentage points on average if a voter is asked to visit the websites of both parties. The estimated causal effect among compliers is larger, showing that those who actually visited the two websites were on average five percentage points more likely to vote than those who did not. Although the 95% Bayesian credible intervals include zero for both ITT effect and CACE, approximately 93% of posterior draws take positive values. The left panel of Figure 2 presents the histograms of posterior simulation draws for one-party and two-party treatments, and graphically illustrates the difference between the two.

We note that the comparison of the two CACE estimates requires some caution because the compliers of different treatments might differ in their characteristics.[28] Although somewhat nuanced, our findings offer support for the hypothesis that additional information increases voter turnout, when voters are exposed to the information of *both* ruling and opposition parties (as opposed to just one party). One possible explanation is that voters can better understand policy differences of the two parties by comparing their policy proposals, and this makes them more likely to vote. Future research should theoretically examine and empirically investigate different causal mechanisms that explain the difference between one-party and two-party treatment effects.

---

[25]These turnout rates for the control group need to be estimated because we do not observe the outcome variable for some voters and the compliance status is not known for every voter in the control group.

[26]It is also possible that the self-reported turnout is biased (e.g., Burden 2000; Silver, Anderson, and Abramson 1986). However, the magnitude of bias in the estimated causal effects due to self-reporting is minimal unless the degree of misreporting is affected by the actual treatment or the treatment assignment.

[27]The estimated CACE for visiting the LDP website is somewhat larger than that of viewing the DPJ website, though the uncertainty estimates become even larger due to small sample size.

[28]This problem is severe when the proportion of noncompliers is large (Cheng and Small 2006; Imai 2005). The comparison of two estimated ITT effects, on the other hand, is straightforward.

TABLE 2   **Estimated Causal Effects of Policy Information on Voter Turnout for One-Party and Two-Party Treatments**

	Summary of Posterior Distributions			
	Mean	s.d.	2.5%	97.5%
One-party treatment				
Intention-to-treat (ITT) effect	0.008	0.021	−0.033	0.051
Complier average causal effect (CACE)	0.011	0.029	−0.044	0.068
Fraction of compliers	0.751	0.007	0.736	0.764
Turnout for the control group				
Compliers	0.706	0.023	0.661	0.752
Noncompliers	0.675	0.057	0.561	0.782
All	0.698	0.009	0.680	0.715
Two-party treatment				
Intention-to-treat (ITT) effect	0.033	0.022	−0.011	0.076
Complier average causal effect (CACE)	0.046	0.032	−0.016	0.108
Fraction of compliers	0.704	0.012	0.680	0.726
Turnout for the control group				
Compliers	0.683	0.027	0.633	0.737
Noncompliers	0.729	0.054	0.618	0.828
All	0.697	0.009	0.680	0.715

Notes: The figures represent the numerical summaries of posterior distributions for each quantity of interest separately for the one-party treatment and two-party treatment conditions: mean, standard deviation, and 95% credible interval. The estimated turnout rates for the control group are shown separately for compliers, noncompliers, and the entire sample, as the baseline turnout of no exposure to the designated party website(s). Each model produces slightly different estimates of turnout rates for the control group. Moreover, compliers are defined separately for each treatment.

FIGURE 2   **Histograms of Posterior Simulation Draws for the Estimated Complier Average Causal Effects (CACE) on Turnout**
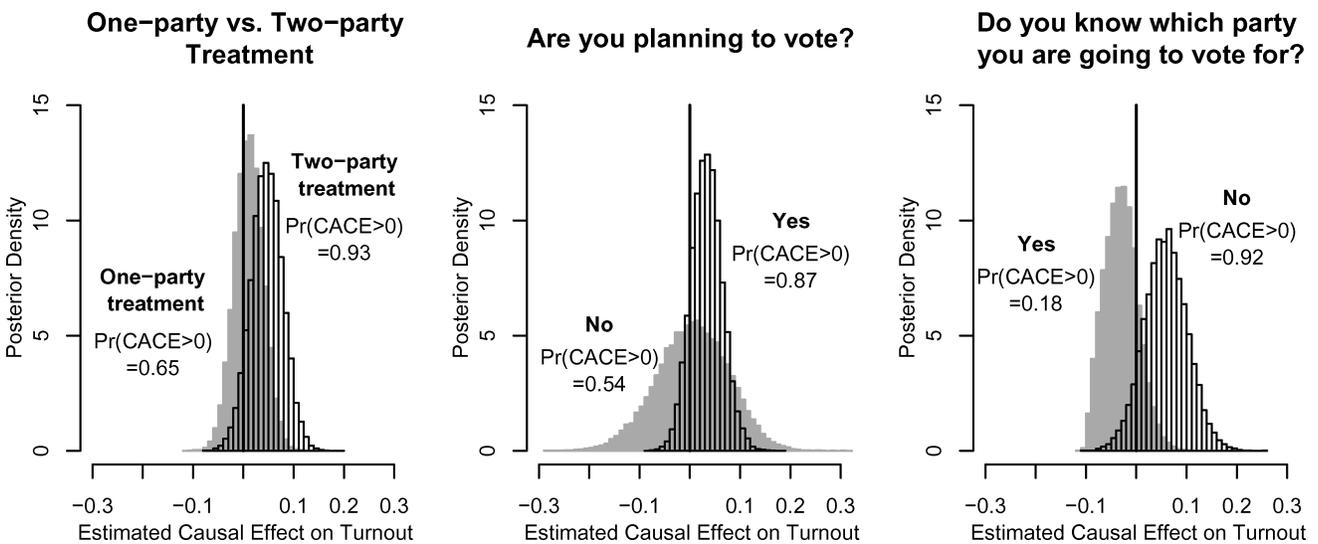
Notes: The left panel compares the one-party treatment effect (shaded histogram) with the two-party treatment effect (unshaded histogram). The middle panel compares CACE using two subgroups—those voters who were planning to vote (unshaded) and those who were not (shaded). The right panel compares the causal effects for another set of two subgroups—those who knew which party they were going to vote for (shaded) and those who did not (unshaded). The vertical lines represent the zero causal effect. The posterior probability that CACE exceeds zero is also shown for each subgroup.

One may argue that the estimated positive effects merely reflect the participation in the survey and are not attributable to the exposure to policy information. To investigate this possibility within our experiment, we conduct sensitivity analyses by redefining those voters who have logged on to the preelection survey as the treated units (see also the previous section). This definition includes the additional 133 voters who started the survey but did not visit the designated website. As expected, we find that the resulting estimated causal effects are smaller by approximately 25%. For the two-party treatment effect, for example, we estimate three percentage points for the ITT effect (posterior standard deviation is 2) and four percentage points for CACE (posterior standard deviation is 3) on average. This suggests that visiting the designated party website(s), rather than participating in the survey, increases turnout. Moreover, regardless of definitions of the actual treatment receipt, the estimated effect of the one-party treatment is smaller than that of the two-party treatment, indicating that different quantities of policy information yield varying degrees of increase in voter turnout. We emphasize that our ability to conduct this kind of detailed analyses depends on the precise measurement of actual treatment receipt we obtained in our preelection survey. This point is made in our third methodological recommendation in the first section.

Results from the Extended Models

We now present the results based on our extended models that address causal heterogeneity.

Results Based on the Simple Approach. Table 3 presents the results for four subgroups that are based on voting intention. Because of the limited sample sizes, we estimate the effect of visiting at least one party's website by pooling the one-party and two-party treatment groups. To test the hypothesis described in the second section, we compare those voters who were planning to vote with those who were undecided or not planning to vote, based on one of the pretreatment variables used to define randomized blocks.[29] The results indicate that visiting at least one designated website increases turnout by three percentage points (from 86%) on average among those who said they were planning to vote. We find very little effect for those who said they were undecided or not planning to vote. While the small sample size makes the finding inconclusive, our treatment had little effect

[29]Because of the limited sample sizes, we pool undecided voters and those who were not planning to vote.

on those who did not have a strong intention to vote in the first place. We consider the effect size quite large for those voters who were planning to vote given that the baseline turnout without exposure to the party website is estimated to be greater than 85%. The middle panel of Figure 2 shows the histograms of posterior distributions of CACE for the two subgroups. About 87% of posterior draws take positive values for the subgroup of those who were planning to vote.

We also compare the group of voters who knew which party they were going to vote for and those who did not know two weeks prior to the election. According to the Downsian hypothesis described in the second section, the uncertain voters are most susceptible to the new information. The results indicate that the information raises turnout by six percentage points on average among those who did not know which party to vote for in the election. Again, although a relatively large posterior standard deviation prevents us from drawing a definitive conclusion, the right panel of Figure 2 shows that 92% of posterior draws are positive for this subgroup. In contrast, visiting at least one designated party website has a slightly negative effect on voters who knew which party to vote for in the election.

Results Based on the General Approach. We first assess the performance of the generalized model by conducting a small simulation study. We generate one pretreatment covariate, X_i, by independently sampling it from $N(0, 1)$ for $i = 1, \ldots, 1000$, and generate the complier indicator variable, C_i, using the compliance model in equation (6) with the true values of coefficients $\xi = (1, 2)$ for the intercept and X_i. Next, we randomly assign the treatment by setting $Z_i = 1$ for 500 randomly selected observations and setting $Z_i = 0$ for the rest. We further assume that the heterogeneity variable, D_i, describing the levels of received treatment, follows $N(X_i, 1)$ for $i = 1, \ldots, 500$, so that X_i and D_i are correlated. Finally, the model in equation (9) is used to generate the outcome variable, Y_i, for each observation with the true value of coefficients $\alpha(D_i) = \cos(\pi D_i)$, $\beta = -1$, $\gamma = (1, 0.5)$. We use the cosine curve to create a complex nonlinear function and hence a difficult test of the proposed model.

Given this simulated data set, we fit our model using the MCMC algorithm. We set $v_0 = s_0^2 = 1$ and run 20,000 iterations. The left panel of Figure 3 summarizes the result of the simulation study by plotting the posterior mean and 95% credible intervals (based on the last 10,000 draws of the resulting Markov chain) along with the true value of the treatment effect parameter, $\alpha(D_i)$, which is represented by the dotted line. The estimates appear to

TABLE 3 Voting Intention and Estimated Causal Effects of Policy Information on Voter Turnout

|	Summary of Posterior Distributions			
	Mean	**s.d.**	**2.5%**	**97.5%**
Planning to vote				
Intention-to-treat (ITT) effect	0.025	0.023	−0.018	0.071
Complier average causal effect (CACE)	0.034	0.031	−0.024	0.095
Fraction of compliers	0.750	0.006	0.738	0.761
Turnout for the control group				
Compliers	0.861	0.022	0.820	0.906
Noncompliers	0.848	0.052	0.738	0.940
All	0.858	0.009	0.838	0.874
Undecided/Not planning to vote				
Intention-to-treat (ITT) effect	0.004	0.049	−0.095	0.097
Complier average causal effect (CACE)	0.006	0.071	−0.137	0.141
Fraction of compliers	0.694	0.009	0.674	0.712
Turnout for the control group				
Compliers	0.315	0.056	0.203	0.424
Noncompliers	0.368	0.089	0.200	0.547
All	0.333	0.022	0.293	0.382
Knew which party they were going to vote for				
Intention-to-treat (ITT) effect	−0.023	0.025	−0.067	0.030
Complier average causal effect (CACE)	−0.030	0.033	−0.088	0.040
Fraction of compliers	0.769	0.008	0.753	0.783
Turnout for the control group				
Compliers	0.929	0.024	0.882	0.973
Noncompliers	0.891	0.062	0.765	1.000
All	0.920	0.010	0.897	0.938
Didn't know which party they were going to vote for				
Intention-to-treat (ITT) effect	0.042	0.031	−0.018	0.103
Complier average causal effect (CACE)	0.059	0.043	−0.026	0.145
Fraction of compliers	0.713	0.006	0.700	0.725
Turnout for the control group				
Compliers	0.555	0.033	0.489	0.621
Noncompliers	0.605	0.066	0.474	0.730
All	0.570	0.013	0.543	0.594

Notes: The figures represent the numerical summaries of posterior distributions for each quantity of interest separately for each subgroup of the sample: mean, standard deviation, and 95% credible interval. The estimated turnout rates for the control group within each subsample are shown separately for compliers, noncompliers, and the entire subsample, as the baseline turnout of no exposure to at least one designated party website.

recover the main feature of the true functional form well in areas where data are abundant, but not where they are sparse.

Finally, we apply the generalized model to our Japanese election data. In particular, we model the two-party treatment effect as a smooth function of the time respondents spent browsing the party websites measured in terms of log seconds. This variable measures the time between when a respondent opened the first of the two designated websites and when he or she answered the next question in the survey. Since we did not control how long respondents had to view the websites, there was significant variation. Some spent only several seconds while others spent more. Overall, the median time was 64 seconds while the 25 and 75 percentiles were 29 and 152 seconds. Since D_i is not randomized, we rely on the assumption

FIGURE 3   **The Performance of the Generalized Model Based on the Simulated Data
(Left Panel) and the Japanese Election Data (Right Panel)**
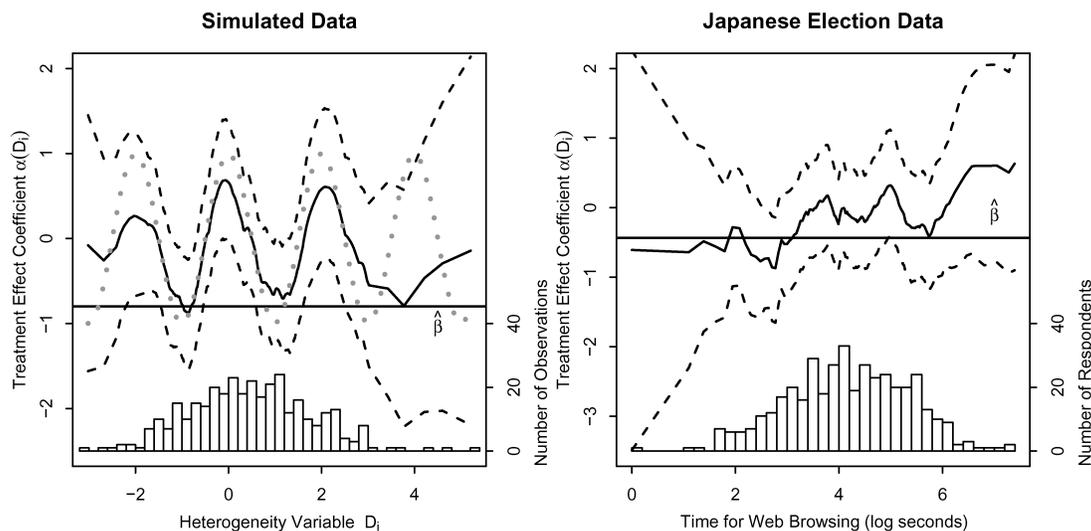
Notes: In the simulation, the true treatment effect function (which is represented by the dotted line in the left panel) is set to $\alpha(D_i) = \cos(\pi D_i)$ where D_i is generated from the normal distribution with mean X_i and unit variance. In the analysis of the Japanese election data, the model estimates the two-party treatment effect as a smooth function of the time respondents spent browsing the designated party websites. In both panels, the solid line represents the posterior mean of the coefficient for the treatment effect $\widehat{\alpha(D_i)}$, and the dashed lines represent the 95% credible interval. The histograms represent the empirical distributions of the heterogeneity variables for the treated units. The solid horizontal lines, denoted by $\hat{\beta}$, represent the posterior mean of the intercept for compliers without treatment. The treatment effect for each observation is a function of $\alpha(D_i) - \beta$.

described in the previous section, by including all of the covariates we used for the basic model. We also include an additional indicator variable in the outcome model for 13 treated compliers for whom we failed to obtain D_i.[30] We also use the same prior specification as before.

The right panel of Figure 3 presents the results based on the last 10,000 draws of the Markov chain after discarding the initial set of 10,000 draws. While the credible interval is wide as is often the case in non/semiparametric models like ours, the graph shows a general upward trend up until around 5 log seconds or 2.5 minutes (after that point, the data are so sparse that the estimate may not be reliable), suggesting that until that point the more time respondents spent viewing the party websites, the higher is the estimated treatment effect.[31] We found little effect

for those who spent less than a total of 30 seconds (or approximately 3.4 on the log scale) on the two party websites. We have argued that the estimated causal effects in our experiment reflect the exposure to the information rather than merely the participation in the survey. The analysis presented here offers some evidence supporting our claim.

Although the generalized model does not assume homogeneity of treatment effects, the model also yields the estimated *average* ITT effect and CACE, averaged over those in the treatment assignment group.[32] It is important to note that these estimated average causal effects remain valid, even if the assumption about the treatment heterogeneity is violated. Indeed, the resulting estimates are quite similar to those presented in Table 2. The posterior means are 3.5 and 4.9 percentage points (with posterior standard deviations of 2.3 and 3.3) for the ITT effect and CACE, respectively.

---

[30]These respondents visited the designated websites more than once, and in such instances D_i was not recorded due to unexpected technical limitations.

[31]If desired, the estimated curve can be made smoother by taking higher-order differences (Koop and Poirier 2004). Alternatively, given the results presented here, one can model based on the linear or quadratic function.

[32]To obtain the ITT effect and CACE for the entire sample, one also needs to model D_i. Since Z_i is randomly assigned, the resulting estimates should not be systematically different, though uncertainty estimates will be smaller.

Concluding Remarks

In this article, we showed how to design and analyze randomized experiments. To illustrate our methods, we designed and analyzed our Japanese election experiment to test Downsian hypotheses about the relationship between information and voter turnout. First, by measuring a number of important pretreatment covariates, researchers can make modeling assumptions more plausible, conduct efficient randomization, and explore causal heterogeneity in treatment effects. Second, in order to ensure efficient randomization, we recommend randomization schemes such as randomized-block and matched-pair designs. Third, the precise measurement of treatment received is essential for accurate interpretations of estimated causal effects and a better understanding of underlying causal processes. Fourth, the statistical method used in this article can overcome the problems of noncompliance and nonresponse simultaneously, which are frequently encountered in randomized experiments. We also developed new ways to model causal heterogeneity in randomized experiments and show the flexibility and potential applicability of the methodological framework we advocate. These and other methods for analyzing randomized experiments are publicly available as an R package, *experiment*, to researchers. Finally, our four methodological recommendations are applicable to other experimental settings, including non-Internet experiments. We hope that our Japanese election experiment serves as a methodological template for future causal inquiry with randomized experiments.

Appendix: Computational Details

This appendix gives details of the Gibbs sampling algorithm that we use to fit the model proposed earlier. The algorithm starts with the initial value for parameters ($\alpha^{(0)}$, $\beta^{(0)}$, $\gamma^{(0)}$, $\xi^{(0)}$) and missing data ($C^{(0)}$, $Y^{(0)}$). The compliance status for the units in the treatment assignment group is known; i.e., $C_i^{(0)} = C_i = T_i$ for units with $Z_i = 1$. Similarly, we set $Y_i^{(0)} = Y_i$ for units with $R_i = 1$. We then proceed via the following steps at iteration t.

Step 1: Sample the binary compliance status $C_i^{(t)}$ independently for each i with $Z_i = 0$ from the Bernoulli distribution with probability,

$$\frac{\lambda_i[Y_i \omega_i + (1 - Y_i)(1 - \omega_i)]}{\lambda_i[Y_i \omega_i + (1 - Y_i)(1 - \omega_i)] + (1 - \lambda_i)[Y_i \pi_i + (1 - Y_i)(1 - \pi_i)]},$$

where $\lambda_i = \Phi(X_i^\top \xi^{(t-1)})$, $\pi_i = \Phi(X_i^\top \gamma^{(t-1)})$, and $\omega_i = \Phi(\beta^{(t-1)} + X_i^\top \gamma^{(t-1)})$. For units with $Z_i = 1$, set $C_i^{(t)} = C_i^{(t-1)}$.

Step 2: Impute the missing values of the outcome variable for each i with $R_i = 0$ by sampling from the Bernoulli distribution with probability $\Phi(\alpha^{(t-1)} C_i^{(t)} T_i + \beta^{(t-1)} C_i^{(t)}(1 - T_i) + X_i^\top \gamma^{(t-1)})$. For units with $R_i = 1$, set $Y_i^{(t)} = Y_i^{(t-1)}$.

Step 3: Given the updated compliance status $C_i^{(t)}$, perform the Bayesian probit regression for the compliance model using the marginal data augmentation scheme in Section 3.3 of Imai and van Dyk (2005). This gives the new draws of the model parameter, $\xi^{(t)}$.

Step 4: Given the updated outcome variable $Y_i^{(t)}$, perform the Bayesian probit regression for the outcome model again using the marginal data augmentation scheme in Section 3.3 of Imai and van Dyk (2005). This gives the new draws of the model parameters, ($\alpha^{(t)}$, $\beta^{(t)}$, $\gamma^{(t)}$).

References

Alvarez, R. Michael. 1998. *Information and Elections: Revised to Include the 1996 Presidential Election.* Ann Arbor: University of Michigan Press.

Angrist, Joshua D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal* 114(494): C52–C83.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90(430): 431–42.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables (with Discussion)." *Journal of the American Statistical Association* 91: 444–55.

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 (with Discussion)." *Journal of the American Statistical Association* 98(462): 299–311.

Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40(1): 194–230.

Basinger, Scott J., and Howard Lavine. 2005. "Ambivalence, Information, and Electoral Choice." *American Political Science Review* 99(2): 169–184.

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8(4): 389–98.

Cheng, Jing, and Dylan Small. 2006. "Bounds on Causal Effects in Three-Arm Trials with Noncompliance." *Journal of the Royal Statistical Society, Series B, Methodological* 68(5): 815–96.

Cochran, W., and G. Cox. 1957. *Experimental Designs*. New York: Wiley.

Cox, David R. 1958. *Planning of Experiments*. New York: John Wiley & Sons.

Cox, David R., and Nancy Reid. 2000. *The Theory of the Design of Experiments*. New York: Chapman & Hall.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: HarperCollins.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4): 627–35.

Ferejohn, John A., and James H. Kuklinski, eds. 1990. *Information and Democratic Processes*. Urbana and Chicago: University of Illinois Press.

Frangakis, Constantine E., and Donald B. Rubin. 1999. "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika* 86(2): 365–79.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1): 21–9.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulations Using Multiple Sequences (with Discussion)." *Statistical Science* 7(4): 457–72.

Greevy, Robert, Bo Lu, Jeffrey H. Silber, and Paul Rosenbaum. 2004. "Optimal Multivariate Matching Before Randomization." *Biostatistics* 5(2): 263–75.

Grofman, Bernard, ed. 1993. *Information, Participation, and Choice: An Economic Theory of Democracy in Perspective*. Ann Arbor: University of Michigan Press.

Hill, Jennifer L., Donald B. Rubin, and Neal Thomas. 1999. "The Design of the New York School Choice Scholarship Program Evaluation." In *Research Designs: Inspired by the Work of Donald Campbell*, ed. Leonard Bickman. Thousand Oaks, CA: Sage Publications.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. 2000. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design." *Biostatistics* 1(1): 69–88.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Replication Data for 'Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment.'" hdl: 1902.1/JMFHKLRCXS http://id.thedata .org/hdl%3A1902.1%2FJMFHKLRCXS. Henry A. Murray Research Archive [distributor (DDI)].

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125 (1–2): 241–70.

Imai, Kosuke. 2005. "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99(2): 283–300.

Imai, Kosuke. 2006. Statistical Analysis of Randomized Experiments with Nonignorable Missing Binary Outcomes. Technical report. Princeton University.

Imai, Kosuke. 2007. "experiment: R package for designing and analyzing randomized experiments." Available at The Comprehensive R Archive Network (CRAN). http://cran .r-project.org.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467): 854–66.

Imai, Kosuke, and David A. van Dyk. 2005. "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation." *Journal of Econometrics* 124(2): 311–34.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2007. Misunderstandings among Experimentalists and Observationalists: Balance Test Fallacies in Causal Inference. Technical report. Princeton University.

Imbens, Guido W., and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25(1): 305–27.

Koop, Gary, and Dale J. Poirier. 2004. "Bayesian Variants of Some Classical Semiparametric Regression Techniques." *Journal of Econometrics* 123(2): 259–82.

Lassen, David Dreyer. 2005. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." *American Journal of Political Science* 49(1): 103–18.

Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Matsumoto, Masao. 2006. "The End of the Independent Voters: A Change of Party Identification (in Japanese)." *Senkyo Kenkyu* 21: 39–50.

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5: 31–61.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Prior, Markus. 2005. "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout." *American Journal of Political Science* 49(3): 577–92.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Richardson, Bradley M. 1997. *Japanese Democracy*. New Haven: Yale University Press.

Sekhon, Jasjeet. 2004. "The Varying Role of Voter Information Across Democratic Societies." Working Paper.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80(2): 613–24.

Tanaka, Aiji. 1997. "Attitudinal Structure of Independent Voters: Rethinking Measurement and Conceptualization of Partisanship (in Japanese)." *Leviathan* 20: 101–29.

Weisberg, Herbert F., and Aiji Tanaka. 2001. "Change in the Spatial Dimensions of Party Conflict: The Case of Japan in the 1990s." *Political Behavior* 23(1): 75–101.

Yatchew, Adonis. 1998. "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature* 36: 669–721.