Randomization Inference With Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election

Daniel E. HO and Kosuke IMAI

Since the 2000 U.S. Presidential election, social scientists have rediscovered a long tradition of research examining the effects of ballot format on voting. Using a new dataset collected by *The New York Times*, we investigate the causal effect of being listed on the first ballot page in the 2003 California gubernatorial recall election. California law mandates a unique randomization procedure of ballot order that, when appropriately modeled, can be used to approximate a classical randomized experiment in a real world setting. We apply randomization inference based on Fisher's exact test, which directly incorporates the exact randomization procedure and yields accurate nonparametric confidence intervals. Our results suggest that being listed on the first ballot page causes a statistically significant increase in vote shares for more than 40% of the minor candidates, whereas there is no significant effect for the top two candidates. We also investigate how randomization inference differs from conventional estimators that do not fully incorporate California's complex treatment assignment mechanism. The results indicate appreciable differences between the two approaches.

KEY WORDS: Causal inference; Fisher's exact test; Nonparametric inference; Permutation test; Political science; Treatment effect.

1. INTRODUCTION

In the 2000 U.S. national election, George W. Bush became President by winning 537 more votes than Al Gore in Florida. Not only did this unusually close election appear to challenge theoretical and empirical propositions that individual voters are rarely decisive (e.g., Riker and Ordeshook 1968; Aldrich 1993; Gelman, King, and Boscardin 1998), but the election also served as a reminder that the manner in which elections are administered can change outcomes. Indeed, the 2000 election spawned a host of scholarly and official investigations into the causal effects of various administrative factors on election outcomes. These factors include the butterfly ballot (Wand et al. 2001), voting equipment (U.S. General Accounting Office 2001; Tomz and Van Houweling 2003), overseas absentee ballots (Imai and King 2004), undervotes (Hansen 2003), and the ballot order of candidates (Krosnick, Miller, and Tichy 2003; Ho and Imai 2004). The election debacle of *Bush v. Gore* also prompted election reform across the United States. Congress authorized nearly $4 billion for voting reform with the Help America Vote Act in 2002 alone.

Whereas the 2000 election highlighted the importance of election administration and ballot format in particular, legal scholars, political scientists, and psychologists have long been interested in examining the causal effects of ballot format on election outcomes (e.g., Gold 1952; Bain and Hecock 1957; Scott 1972; Darcy 1986; Darcy and McAllister 1990; Miller and Krosnick 1998). But studies typically use observational data, in which the resulting estimates of causal effects are subject to potential confounding factors, and laboratory experiments, in which results may lack external validity. We address these shortcomings by analyzing a randomized natural experiment. Because randomized experiments are difficult to conduct in real elections for ethical and practical reasons, natural experiments provide rare opportunities to make causal inferences with both internal and external validity.

In particular, we study the causal effect of the page placement of candidates in the 2003 California recall election using a unique dataset that was collected by *The New York Times* (Kershaw 2003). Since 1975, California law has mandated that the Secretary of State draw a random alphabet for each election to determine the order of candidates for the first assembly district [California Election Code § 13112 (2003)]. California law further requires that the candidate order be systematically rotated throughout the remaining assembly districts. We exploit this randomization-rotation procedure to estimate the causal effect of being listed on the first ballot page on a candidate's vote share. This question is important from two perspectives. First, from a behavioral voting perspective, our study investigates whether voters are able to act as if they are fully informed (e.g., Bartels 1996; Forsythe, Myerson, Rietz, and Weber 1993). Second, from a policy making perspective, ballot design is seen as central to electoral fairness and design (e.g., Garrett 2004; the "Making Your Vote Count" series of *The New York Times* editorials on voting fairness, 2004).

The analysis of the 2003 California recall election also poses unique statistical challenges. First, treatment assignment is randomized but in an unconventional way. The units of randomization are the alphabet letters rather than the candidates, and the randomization is followed by systematic rotation of the candidate order through 80 nonrandomly ordered assembly districts. Second, the 58 counties in California each print unique ballots, and an assembly district may contain more than one county and/or only a part of a county. Third, the unusually high level of media attention and low threshold of ballot access led

Daniel E. Ho is Assistant Professor of Law, Stanford Law School, Stanford, CA 94305 (E-mail: *daniel.e.ho@gmail.com*). Kosuke Imai is Assistant Professor, Department of Politics, Princeton University, Princeton, NJ 08544 (E-mail: *kimai@princeton.edu*). The authors are grateful to Don Rubin for helpful comments at an earlier stage of this project. They thank Chris Adolph, Jim Alt, Henry Brady, Andy Gelman, Marshall Joffe, Gary King, Jon Krosnick, Kevin Quinn, Adrian Raftery, Paul Rosenbaum, Jas Sekhon, Liz Stuart, seminar participants at Columbia University and the University of Washington, three anonymous referees, the associate editor, and the editor for helpful comments. An earlier version of this article was presented at the 2004 Annual Summer Meeting of the Society for Political Methodology and the 2005 Annual Meeting of the Midwest Political Science Association. The authors also thank Archie Tse of *The New York Times* and Evelyn Mendez of the California Secretary of State's Office for data. Research support was provided by the National Science Foundation (SES-0550873), as well as the Committee on Research in the Humanities and Social Sciences at Princeton University, Institute for Quantitative Social Science, the Project on Justice, Welfare and Economics, the Center for American Political Studies at Harvard University, and the Center for Law, Economics, and Organization at Yale University.

Journal of the American Statistical Association
September 2006, Vol. 101, No. 475, Applications and Case Studies
DOI 10.1198/016214505000001258

to an unprecedented total of 135 candidates, from Hollywood actor Arnold Schwarzenegger, who eventually won the election, to child television star Gary Coleman. Challenges arise primarily out of the fact that the randomization-rotation procedure was designed by policy makers for ease of implementation, not for ease of statistical estimation. Therefore, careful statistical analysis is required to draw valid causal inferences. Given the peculiarity of the recall election, we limit ourselves to in-sample inferences; a comprehensive analysis of other California elections appears elsewhere (Ho and Imai 2004).

To address these challenges, we apply randomization inference, which was originally developed by Fisher (1935) and later extended by others (see, e.g., Cox and Reid 2000; Rosenbaum 2002c). More recently, randomization inference has been applied to observational studies (e.g., Rosenbaum 2002a,b) and instrumental variables (e.g., Rosenbaum 1996; Greevy, Silber, Cnaan, and Rosenbaum 2004; Imbens and Rosenbaum 2005). Our application illustrates that randomization inference may be applied to various treatment assignment mechanisms, such as the alphabetical randomization of the 1970 Vietnam draft lottery (Starr 1997; Angrist 1990), randomization-rotation commonly applied to reduce survey question order effects in psychology (Shaughnessy, Zechmeister, and Zechmeister 2002, chap. 7) and sequential randomization on covariates in clinical trials (Pocock and Simon 1975).

We first use an extension of Fisher's exact test to examine the sharp null hypothesis of no ballot page effect. The result suggests that being placed on the first page of ballot was associated with a significant increase in vote shares for more than 40% of the candidates. In contrast to the results based on conventional estimators, we find that (a) page placement does not decrease vote shares and (b) there are no significant effects of page placement for the top two candidates. These findings are consistent with the results of Ho and Imai (2004), who analyzed the causal effect of being placed first on the ballot (rather than being placed on the first page) in other California general elections. Next, we invert Fisher's exact test to obtain nonparametric confidence intervals for ballot page effects.

The article is organized as follows. In Section 2 we explain the randomization-rotation procedure mandated by California law and describe our dataset of the 2003 California recall election. In Section 3 we place our analysis in a statistical framework of causal inference and explain how Fisher's exact test can be extended to conduct distribution-free hypothesis testing about causal effects. We also show how to obtain nonparametric confidence intervals of quantities of interest by inverting the test. In Section 4 we present the results of our analysis based on randomization inference, conduct sensitivity analyses, and compare randomization inference with conventional estimators that do not fully incorporate the treatment assignment mechanism. We conclude in Section 5.

2. RANDOMIZATION–ROTATION PROCEDURE AND RECALL ELECTION DATA

In this section we briefly explain the randomization-rotation procedure used for California statewide elections. We also describe our dataset of California ballots, which was originally collected by *The New York Times*. To supplement this data, we also collected official election returns, voter registration data, and Census data.

2.1 California Alphabet Lottery

Until 1975, incumbents appeared first on the ballot in most California statewide elections. But then the California Supreme Court ruled in *Gould v. Grubb*, 14 Cal. 3d 661 (1975) that listing candidates by incumbency or alphabetical order was unconstitutional (see also Scott 1972). In response, the California legislature enacted a randomization procedure to determine the order of candidates. According to California Elections Code § 13112 (2003), the alphabet lottery works in three steps. First, the Secretary of State randomly draws the letters of the alphabet, so that all 26! possible permutations of the alphabet are equally possible. Second, names of candidates for each statewide office are ordered by this randomized alphabet for the first of 80 assembly districts. Third, candidate names are systematically rotated for each subsequent assembly district. That is, the candidate listed first in a district moves to the last place in the next district, and of all the other candidates move upward by one position. In a typical race with between five and seven candidates, for example, this would ensure that all candidates were listed roughly an equal number of times.

For the 2003 recall election, the actual randomized alphabet was

R W Q O J M V A H B S G Z

X N T C I E K U P D Y F L.

Based on this randomized alphabet, the ballot order in the first assembly district was determined, starting from Robinson, Roscoe, Ramirez, and so on and proceeding to Lewis and Leonard. This candidate order was then rotated throughout the remaining assembly districts.

Ho and Imai (2004) first used the California alphabet lottery to estimate the causal effect of ballot order on candidates for statewide elections from 1978 to 2002. Their statistical tests confirmed that the resulting alphabets from the California alphabet lottery were indeed random (see also Fig. 3 for tests to show the complete randomization of page placement in the recall data). Ho and Imai (2004) also identified several statistical challenges. Most importantly, the randomization-rotation procedure poses difficulties in identifying the variance of conventional estimators such as a difference-in-means estimator and a linear regression estimator. Ho and Imai (2004) pointed out that a similar situation arises in systematic sampling in surveys; the fact that randomization in systematic sampling occurs only once makes identifying the variance difficult without making distributional assumptions about the population order (see, e.g., Cochran 1977, chap. 8; Wolter 1984).

In our application, this variance identification problem is exacerbated because the population order of the California districts is nonrandom, its distribution is unknown, and the number of districts is small relative to the number of candidates. Moreover, the unit of randomization is not the page position of a candidate, but rather the alphabet, and ballot pages vary across counties mainly because each county uses different voting equipment. This leads to unequal probability treatment assignment that may be confounded, while making the identification of standard variance estimators even more difficult. Ho and Imai (2004) found that in typical California elections these challenges are not severe when analyzing the effect of

being listed first on the (one-page) ballot (rather than the effect of being placed on the first page of the multipage ballot). This is because in contrast to the recall election, the number of candidates in typical elections is small (generally around five) relative to the number of districts, and so the randomization-rotation procedure leads to a good balance of observed district characteristics and roughly equal probability assignment.

2.2 Recall Election Dataset

Our 2003 California recall election data consist of the ballot page placement data collected by *The New York Times*, supplemented by 2000 Census data as well as official election results and registration data obtained from the California Secretary of State. The dataset comprises geographical units defined by counties and assembly districts. Each of the 58 counties uses a different ballot format with varying numbers of pages, and the candidate order differs by each of 80 assembly districts. Therefore, the page on which the name of each candidate appears depends on both (a) the *ballot format* determined by each county registrar, which is not randomized, and (b) the *ballot order* in each of the 80 assembly districts, which is determined by randomized alphabet and systematic rotation.

For example, Del Norte County and Humboldt County both belong to the first assembly district. But while Humboldt County uses a one-page ballot, listing all 135 candidates together on a single page, Del Norte County uses a five-page ballot, listing only 23 candidates on the first page. Butte County is split into two assembly districts; Schwarzenegger is listed on the fourth page in the second assembly district but on the third page in the third assembly district. In total, we have 158 unique assembly district–counties, out of which 121 units have ballots with more than one page. (For simplicity, we call these geographical units "districts" throughout the remainder of the article.) These 121 districts serve as the units of our analysis. We exclude the districts with one-page ballots, because for these districts the in-sample causal effect of being listed on the first page cannot be defined.

The New York Times data contain information on the page placement of 135 candidates in each district. The data from the California Secretary of State Office provide certified election returns of all candidates as well as party registration rates (i.e., the number of registered Republican or Democratic voters divided by the total number of registered voters) in each district. One limitation of *The New York Times* dataset is that it does not contain information about the placement of candidate names *within* each ballot page. Using the 2000 Census, we also collected data on income (mean household wage or salary income in 1999) and gender and racial compositions (proportions of male, whites, Asians, Latinos, and African-Americans for each district, all of which take a value between 0 and 1). These variables are pretreatment covariates because they were measured before the randomization of the treatment (although not necessarily before the selection of ballot format for each district).

Figure 1 summarizes the page placement of each candidate by the number of districts in which the candidate was listed on the first page. We analyzed a total of 121 districts with multipage ballots. The vertical axis lists each of the 135 candidates in order of the randomized alphabet. The horizontal axis represents the number of districts. The dark shading of the horizontal bars indicates the number of districts in which voters

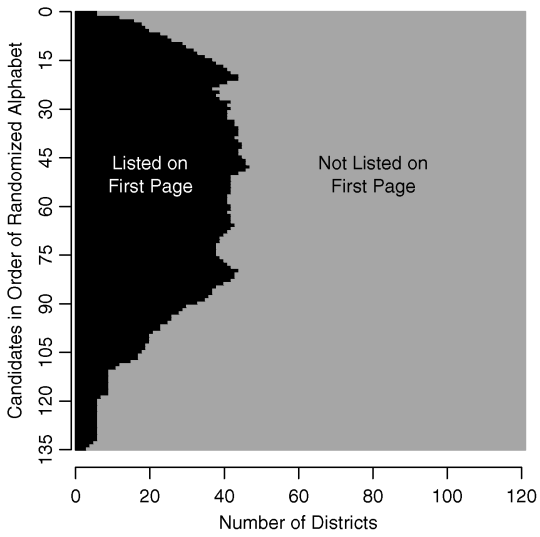

Figure 1. Page Placement for 135 Candidates by the Number of Districts. Candidates are listed in order of the randomized alphabet. For each candidate, the dark shading indicates the total number of districts in which the candidate was listed on the first page. The gray shading indicates the total number of districts in which the candidate was not listed on the first page. Districts with single-page ballot are excluded from the graph as well as from our analysis. The total number of districts with multipage ballots is 121.

observed the candidate on the first page, whereas the gray shading corresponds to the districts in which the candidate was not listed on the first ballot. The figure suggests that the California alphabet lottery does not result in complete randomization of page placement across candidates or across districts. For example, Robinson, who is first in the randomized alphabet and represented by the top horizontal bar as candidate 1, appears on the first page in only 6 multipage districts, whereas Schwarzenegger, who is candidate 74, is listed on the first page in 38 districts. The 25 candidates at the end of the randomized alphabet are all listed on the first page in fewer than 10 districts.

3. METHODOLOGY

In this section we present our approach to estimating ballot page effects. We place our analysis in the formal statistical framework of causal inference based on potential outcomes (e.g., Holland 1986). Within this framework, we describe randomization inference derived from Fisher's exact test (e.g., Rosenbaum 2002c). The key insight is to incorporate the exact randomization-rotation procedure as a central part of statistical estimation. In addition to testing the sharp null hypothesis of no unit treatment effect, we invert Fisher's exact test to obtain nonparametric confidence intervals. We use a simple extension of the bisection algorithm to conduct such distribution-free inferences.

3.1 Framework of Causal Inference

We conduct our analysis separately for each of the 135 candidates. For each candidate, we observe vote shares for 121 districts with multipage ballots. Let y_i denote the *observed* vote share for the ith district. For each district $i = 1, 2, \ldots, 121$, we define two *potential* outcomes, Y_{1i} and $Y_{0i} \in [0, 1]$. Y_{1i} denotes the potential vote share in district i when the candidate is placed

on the first page of the ballot, whereas Y_{0i} represents the potential vote share in district i when the candidate is not on the first page. We use the indicator variable, $T_i \in \{0, 1\}$, to denote the treatment status in district i. T_i equals 1 if candidate i is listed on the first page and 0 otherwise. Thus the observed outcome variable is a function of the treatment variable and potential outcomes, $Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$. We use upper-case letters to distinguish a random variable Y_i from its realization y_i. The fundamental problem of causal inference is that we observe only one of the two potential outcomes (Holland 1986).

Instead of estimating the effect of being listed on the first page, it may also be possible to analyze the effect of being listed on each page as a multitreatment regime by assuming a constant additive treatment effect (see, e.g., Angrist and Imbens 1995; Imai and van Dyk 2004; Imbens and Rosenbaum 2005). But this assumption is implausible, because the effect of being listed on the first versus the second page, for example, may well differ from the effect of being listed on the fifth versus the sixth page (see, e.g., Ho and Imai 2004). Alternatively, one might dichotomize the treatment for each position, which is a straightforward extension of our approach. Similarly, if the data were available, one might also estimate the vector of potential outcomes for page placement and the ballot order within each page.

Our adopted framework of potential outcomes implicitly makes the following assumption (Cox 1958; Rubin 1990):

Assumption 1 (No interference among units). The potential outcomes of one unit do not depend on the treatment of other units.

In our application, making this assumption implies that potential vote shares of a candidate in one district do not depend on the same candidate's ballot placement in another district. This assumption is likely to hold, because voters usually do not see ballots of other districts and hence are unlikely to be affected by such ballots.

Assumption 1 would be violated if we considered joint vote shares of all 135 candidates, which must sum to unity within each district. To model candidates jointly, however, would imply an extremely large number of missing potential outcomes requiring strong assumptions for identification (e.g., that the treatment effect of one candidate draws votes proportionally from all other candidates). Because such assumptions are implausible in this application, we focus on the estimation of a separate causal effect for each candidate, which makes Assumption 1 more reasonable.

Within this framework of causal inference, we consider in-sample inferences where Y_{1i} and Y_{0i} are assumed to be fixed (but not necessarily observed) quantities. From this perspective, the treatment variable T_i is the only random variable and, as explained later, completely determines the reference distributions of test statistics under the null hypothesis. (Because Y_i is a function of T_i, it is also a random variable.) Given this setup, we define the *unit ballot page effect* (or treatment effect) in the ith district as

$$\tau_i \equiv Y_{1i} - Y_{0i}, \tag{1}$$

which is also a fixed quantity. To make inferences beyond the sample at hand, researchers typically consider a repeated-sampling process and treat Y_{1i} and Y_{0i} as random variables (e.g., Imbens 2004). In this application, however, we confine ourselves to in-sample inferences. Substantively, this means that we investigate only the causal effects in the 2003 recall election. Because an unprecedented number of candidates competed in the recall election and media coverage was unusually high, a population for which to draw out-of-sample inferences may be difficult to define.

3.2 Randomization Inference Using Fisher's Exact Test

When making in-sample causal inferences using Fisher's exact test, we use the null hypothesis about the unit ballot page effect defined in (1). In particular, we hypothesize that the unit treatment effect is zero for all districts,

$$H_0 : \tau_i = 0 \quad \text{for all } i = 1, \ldots, 121, \tag{2}$$

which indicates that (potential) candidate vote shares in each district are the same irrespective of candidate page placement. This null hypothesis said to be *sharp* because it is about the treatment effect of each observation rather than its average over a group of observations.

Under the sharp null hypothesis, all potential outcomes are known exactly. Consider, for example, the units in the treatment group. For these units, we observe one of the *fixed* potential outcomes under the treatment, that is, $y_i = Y_{1i}$, but Y_{0i} is missing. Under the null hypothesis, however, the missing outcome is equal to the observed outcome, $Y_{0i} = y_i$ (similarly, $Y_{1i} = y_i$ for the control units). Given this setup, we formulate the following test statistic, which corresponds to the differences-in-means estimator for the sample average treatment effect:

$$W^{\mathrm{D}}(\mathbf{T}) = \frac{\sum_{i=1}^{121} T_i y_i}{N_1} - \frac{\sum_{i=1}^{121} (1 - T_i) y_i}{N_0}, \tag{3}$$

where $\mathbf{T} = (T_1, T_2, \ldots, T_{121})$, $N_1 = \sum_{i=1}^{121} T_i$, and $N_0 = 121 - N_1$. Alternatively, we also formulate a covariance-adjusted test statistic by linear least squares (Rosenbaum 2002b),

$$W^{\mathrm{L}}(\mathbf{T}) = (\mathbf{T}^\top \mathbf{M} \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{M} \mathbf{y}, \tag{4}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_{121})$, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, and $\mathbf{X}$ is the matrix of the observed pretreatment covariates. We denote the observed value of these statistics as $W^{\mathrm{D}}(\mathbf{t})$ and $W^{\mathrm{L}}(\mathbf{t})$ with the observed treatment status $\mathbf{t} = (t_1, t_2, \ldots, t_{121})$. Other potential test statistics that may be used here include the median test statistic and various rank sum statistics (e.g., Wilcoxon 1945; see Sec. 4.2). We choose $W^{\mathrm{D}}(\mathbf{T})$ and $W^{\mathrm{L}}(\mathbf{T})$ because they correspond to conventional estimators discussed in Section 4.3, where we investigate differences between randomization inference and conventional estimators based on these identical test statistics. The two statistics also represent our scientific quantity of interest, providing an intuitive interpretation.

The notations $W^{\mathrm{D}}(\mathbf{T})$ and $W^{\mathrm{L}}(\mathbf{T})$ emphasize the fact that under the null hypothesis, only the treatment variable $\mathbf{T}$ is random. Therefore, the reference distributions of the test statistics are completely determined by the randomization distribution of $\mathbf{T}$. We assume knowledge of random assignment:

Assumption 2 (Known random assignment). Treatment is randomly assigned by a known mechanism. Formally, $p(T_i | Y_{1i}, Y_{0i}) = p(T_i)$ is known for each i.

In the context of the recall election, this assumption also implies that county page formats are independent of the randomized alphabet. This would be violated if county officials designed ballot pages based on each election's randomized alphabet. Such a scenario is unlikely, because the number of possible ballot pages is driven primarily by the type of voting technology (Kershaw 2003). Because voting technology is exogenous to the randomization (e.g., officials did not change voting technology after observing that Schwarzenegger was randomized to the end of the alphabet), the assumption is likely to hold.

With knowledge of the assignment mechanism, the exact distribution of W under the null hypothesis can be derived. The exact (one-tailed) p-values are then defined by

$$\begin{aligned} p^{\mathrm{D}} &\equiv \Pr\bigl(W^{\mathrm{D}}(\mathbf{T}) \geq W^{\mathrm{D}}(\mathbf{t})\bigr), \\ p^{\mathrm{L}} &\equiv \Pr\bigl(W^{\mathrm{L}}(\mathbf{T}) \geq W^{\mathrm{L}}(\mathbf{t})\bigr). \end{aligned} \tag{5}$$

The null hypothesis is rejected when this p-value is less than a predetermined significance level. In principle, this test is *exact* in the sense that it requires no large-sample approximation. (For computational reasons that we explain later, we approximate the exact distribution of each test statistic with Monte Carlo simulation.) The test is also *distribution-free*, because it does not impose distributional assumptions that are typically invoked to approximate the reference distribution in standard hypothesis testing.

In our application, the California alphabet lottery procedure defines the random treatment assignment and hence determines the exact distribution of the test statistic. As described in Section 2, treatment assignment in each district is determined *systematically* once the alphabet letters are randomized. Directly incorporating the exact randomization procedure may be difficult in standard parametric frameworks. Although there are 135 ways to order a particular candidate in the first district, each order is not equally likely. Moreover, with 135 candidates and 121 districts, systematic rotation does not ensure that each candidate is placed in the same position with equal probability in each district. Indeed, as shown in Section 2.2, there is substantial variation in the page placement of candidates. Accounting for such complications may be difficult, especially when estimating the variances of conventional estimators such as least squares and difference in means. A similar situation arises in systematic sampling methods in the context of survey sampling where single randomization is followed by systematic rotation (e.g., Cochran 1977).

In contrast, Fisher's exact test allows us to directly incorporate the exact randomization procedure, that is, the systematic rotation of the treatment after a random start. In natural experiments, such deviations from simple random assignment may be common (see, e.g., Starr 1997). In the California alphabet lottery, there are $26! \approx 4.0 \times 10^{26}$ ways to order the alphabet letters. Given a particular permutation of alphabet letters, the candidate names are ordered and the treatment within each district, T_i, is assigned deterministically. Because we analyze each candidate separately, this means that alphabet randomization gives different weights to each of 135 ways of ordering a particular candidate in the first district. We can then compute the exact distributions of the test statistics, $W^{\mathrm{D}}(\mathbf{T})$ and $W^{\mathrm{L}}(\mathbf{T})$, by calculating the values that they take given each permutation of alphabet letters.

In this application, the expressions in (5) are difficult to evaluate analytically because of the complex nature of the treatment assignment rule used for California statewide elections. Moreover, a large number of permutations is produced by alphabet randomization. Therefore, we use the following Monte Carlo approximation to compute the p-values:

$$\begin{aligned} p^{\mathrm{D}} &\approx \frac{1}{m}\sum_{j=1}^{m} I\bigl\{W^{\mathrm{D}}\bigl(\mathbf{T}^{(j)}\bigr) \geq W^{\mathrm{D}}(\mathbf{t})\bigr\}, \\ p^{\mathrm{L}} &\approx \frac{1}{m}\sum_{j=1}^{m} I\bigl\{W^{\mathrm{L}}\bigl(\mathbf{T}^{(j)}\bigr) \geq W^{\mathrm{L}}(\mathbf{t})\bigr\}, \end{aligned} \tag{6}$$

where $\mathbf{T}^{(j)}$ is the jth draw of the random variable from its known distribution, $I\{\cdot\}$ is the indicator function, and m is the total number of draws to approximate the distribution. That is, we randomly order alphabet letters and then deterministically obtain the treatment assignment for each district. After testing various values of m, we find that 10,000 is sufficiently large to provide a reliable approximation in our application.

3.3 Nonparametric Confidence Intervals

In-sample randomization inference using Fisher's exact test can be further extended by inverting the test. Test inversion is a standard way to obtain confidence intervals (e.g., Cox and Hinkley 1979). The resulting confidence intervals based on Fisher's exact test are distribution-free and have accurate coverage probabilities.

To invert the test, we first assert a general null hypothesis, under which the unit treatment effect is assumed to be constant across all units in the sample,

$$H_0 : \tau_i = \tau_0 \quad \text{for all } i = 1, \ldots, 121, \tag{7}$$

for some constant τ_0. Under this sharp null hypothesis, missing potential outcomes are known exactly as before. For the units in the treatment group, the missing outcome can be computed by $Y_{0i} = y_i - \tau_0$ under the null hypothesis (similarly, $Y_{1i} = y_i + \tau_0$ for the units in the control group). Given this sharp null hypothesis, we generalize our test statistic, $W^{\mathrm{D}}(\mathbf{T})$, to incorporate arbitrary values of τ_0,

$$W^{\mathrm{D}}_{\tau_0}(\mathbf{T}) = \frac{\sum_{i=1}^{121} T_i\{y_i + (1 - t_i)\tau_0\}}{\sum_{i=1}^{121} T_i} - \frac{\sum_{i=1}^{121}(1 - T_i)(y_i - t_i\tau_0)}{\sum_{i=1}^{121}(1 - T_i)}. \tag{8}$$

The test statistic in (3), which is based on the difference-in-means estimator, is simply a special case of $W^{\mathrm{D}}_{\tau_0}(\mathbf{T})$ when $\tau_0 = 0$. The covariance-adjusted analog is

$$W^{\mathrm{L}}_{\tau_0}(\mathbf{T}) = (\mathbf{T}^{\top}\mathbf{M}\mathbf{T})^{-1}\mathbf{T}^{\top}\mathbf{M}\mathbf{y}^{*}, \tag{9}$$

where each element of $\mathbf{y}^{*}$ is $y_i^{*} = T_i\{y_i + (1 - t_i)\tau_0\} + (1 - T_i)(y_i - t_i\tau_0)$. The test statistic in (4), which is based on the linear least squares regression, can be obtained by setting $\tau_0 = 0$. Finally, we denote the observed values of the test statistics by $W^{\mathrm{D}}_{\tau_0}(\mathbf{t})$ and $W^{\mathrm{L}}_{\tau_0}(\mathbf{t})$. Note that these observed values do not depend on τ_0.

As before, the treatment assignment, T_i, is the only random variable, and everything else is known and fixed under the sharp null hypothesis. Therefore, the distribution of the test statistics under any null value τ_0 is solely determined by that of T_i. Then our level-α hypothesis test (two-tailed) is described by the following decision rule: accept H_0 if

$$t \in A^{\mathrm{D}}_{\alpha}(\tau_0) = \left\{ \mathbf{u} : \frac{\alpha}{2} \leq \Pr\left(W^{\mathrm{D}}_{\tau_0}(\mathbf{T}) \geq W^{\mathrm{D}}_{\tau_0}(\mathbf{u})\right) \leq 1 - \frac{\alpha}{2} \right\}, \quad (10)$$

and reject H_0 otherwise, where $A^{\mathrm{D}}_{\alpha}(\tau_0)$ denotes the acceptance region of the test for the difference-in-means test statistic, $W^{\mathrm{D}}_{\tau_0}(\mathbf{T})$. For the covariance-adjusted test statistic, $W^{\mathrm{L}}_{\tau_0}(\mathbf{T})$, the acceptance region, $A^{\mathrm{L}}_{\alpha}(\tau_0)$, can be defined in the same way.

Within this general setup, we obtain the $(1-\alpha)$ confidence intervals of τ by inverting the test. Here we describe our method using the test statistic $W^{\mathrm{D}}_{\tau_0}(\mathbf{T})$; the same procedure can be used for $W^{\mathrm{L}}_{\tau_0}(\mathbf{T})$. First, we note that the $(1-\alpha)$ confidence set is given by $C_{\alpha}(\mathbf{t}) = \{\tau : \mathbf{t} \in A^{\mathrm{D}}_{\alpha}(\tau)\}$. Second, we define the $(1-\alpha)$ confidence interval as the shortest interval that includes the $(1-\alpha)$ confidence set. Then we can compute the confidence interval by identifying the upper and lower bounds, $\tau_{\mathrm{L}} = \sup_{\tau} A^{\mathrm{D}}_{\alpha}(\tau)$ and $\tau_{\mathrm{U}} = \inf_{\tau} A^{\mathrm{D}}_{\alpha}(\tau)$. To obtain upper and lower bounds of confidence intervals, we use a simple extension of the bisection algorithm (e.g., Lange 1999), details of which are given in the Appendix.

Finally, although this application of Fisher's exact test allows for continuous outcomes and a large variety of treatment assignment mechanisms, the exact test is more commonly used in testing the equality of two independent binomial proportions. However, the fundamental idea of the generalized application already existed in Fisher (1935) and Kempthorne (1952).

4. RESULTS FROM THE 2003 CALIFORNIA RECALL ELECTION

In this section we analyze our dataset of the 2003 California recall election. We first present the results based on randomization inference, which directly incorporates the known treatment assignment mechanism. We also explore the possibilities of using other test statistics and relaxing the constant additive treatment effect assumption. Finally, we compare these results with conventional estimators, which do not fully incorporate the randomization procedure, and empirically examine the consequences of ignoring the assignment mechanism.

4.1 Randomization Inference

We first test the sharp null hypothesis of no unit treatment effect as described in Section 3.2. Figure 2 presents (Monte Carlo approximated) one-tailed p-values from Fisher's exact test for each candidate, using $W^{\mathrm{D}}(\mathbf{T})$ as the test statistic. Candidates are ordered by the size of their p-values. Here we follow others (e.g., Rosenbaum 2002c) and use the one-tailed p-values by hypothesizing that being listed on the first page does not decrease a candidate's vote share. Under the null hypothesis of no unit treatment effects, these p-values are expected to be uniformly distributed (i.e., roughly following the diagonal line of Fig. 2). Instead, p-values are very small for a disproportionate number of candidates, all of whom are minor candidates. For 59 out of 135 candidates, we reject the null hypothesis at the α level of .1. (Because for each candidate the probability of rejecting the null hypothesis when the null hypothesis is true is .1, the test is expected to reject about 14 candidates by chance even when there is no ballot effect for all candidates.) Consistent with the main results of Ho and Imai (2004), we find that none of major candidates exhibits statistically significant page effects.

To further illustrate these results, Figure 3 presents the p-values from Fisher's exact test in the same manner as in Figure 2, except that we test the null hypothesis of no effect on pretreatment variables rather than on candidates' vote shares. We present the results for three covariates (number of registered voters, proportion male, and Republican vote share in the 2002 gubernatorial election). Because the treatment is randomized, we would expect no significant page effects on these variables, which were measured before randomization of the treatment. Under the sharp null hypothesis of no page effects, the p-values are expected to be distributed uniformly, roughly following the diagonal line. Figure 3 shows that, as expected, these pretreatment variables are not affected by candidates' page placement. In contrast, Figure 2 shows that for a large number of minor candidates, page placement has a statistically significant effect on vote shares.

Next, we invert Fisher's exact test to obtain nonparametric confidence intervals. Figure 4 presents 90%, 80%, 70%, and 60% confidence intervals obtained from the inversion of (10). For 16 candidates, the 90% confidence intervals include the entire sample space, indicating that the data contain no information about the effects of page placement for these candidates. (For one candidate, only the upper bound is defined.) The reason for this is that these candidates appeared on the first page of multipage ballots in a very small number of districts because they were listed near the end of the candidate order for the first district (see Fig. 1) and/or are minor candidates who received no votes in many districts. This demonstrates one important virtue of the nonparametric approach. When the data are uninformative, randomization inference cannot identify a confidence interval, whereas a parametric approach may still provide estimates. Imbens and Rosenbaum (2005) made an analogous point, arguing that this is a primary reason for using nonparametric confidence intervals for weak instrumental variables.

Substantively, our randomization-based confidence intervals suggest that major candidates are not significantly affected by page placement. The 90% nonparametric confidence interval for the Democratic nominee, Bustamante, for example, is $(-.20, .20)$. Similarly, for Schwarzenegger, Huffington, and McClintock, randomization inferences detect no significant effects. Moreover, being listed on the first page rarely hurts candidates, confirming our hypothesis about the one-tailed tests shown in Figure 2. Randomization inference yields significantly negative effects for only four candidates. The possibility of losing votes due to early ballot placement has been a point of debate among political scientists, with Miller and Krosnick (1998) asserting that candidates might exhibit "recency effects" of gaining votes when listed later in the ballot. In contrast, our randomization inference suggests that in the recall election, candidates did not lose votes from being listed on the first page. This finding is consistent with the results of Ho and Imai (2004) for other California elections.

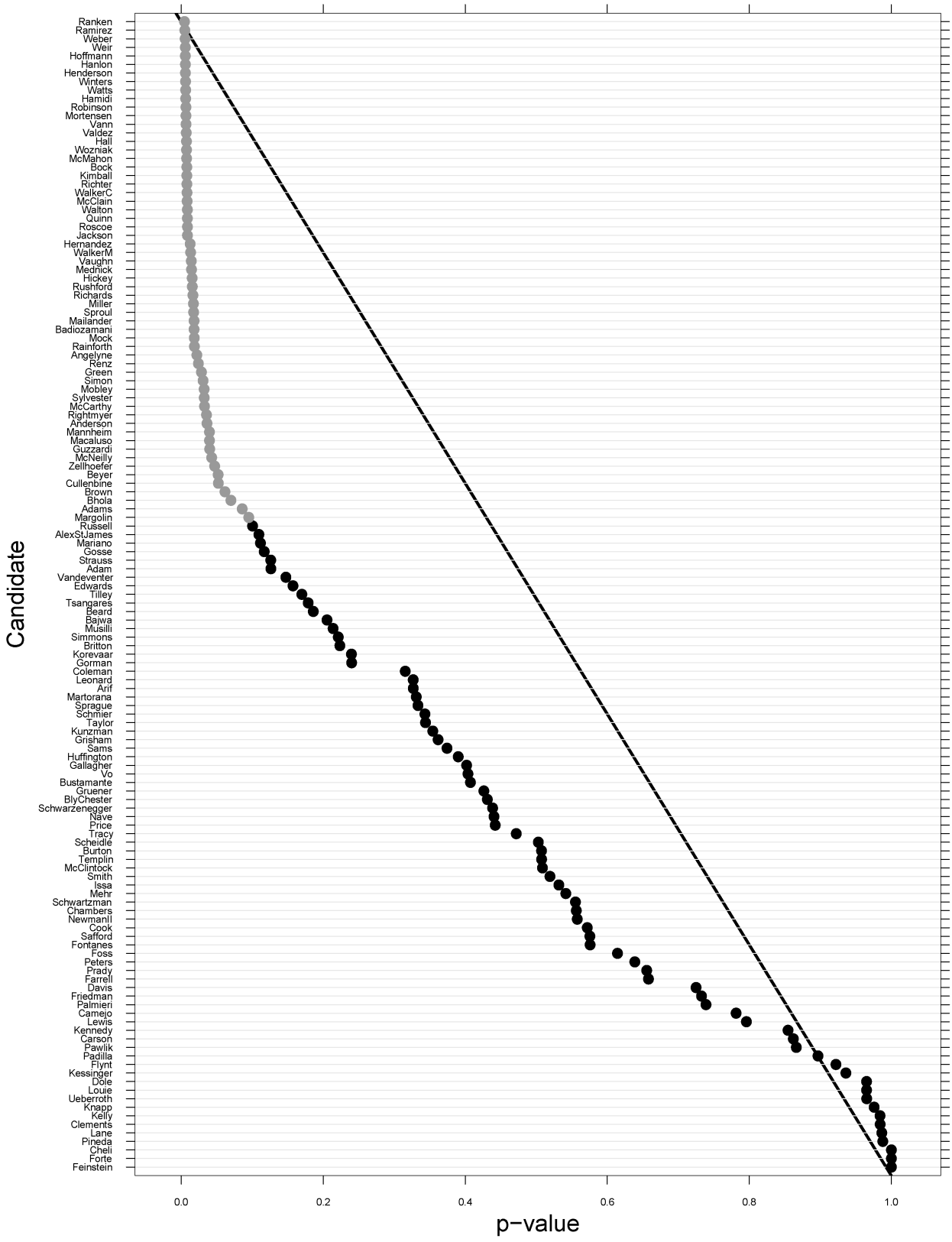

Figure 2. One-Tailed p-Values From Randomization Inference for All 135 Candidates, Arranged in Order of Magnitude. Gray dots denote the p-values that are $<.1$. Under the sharp null hypothesis of no ballot page effects, the p-values are expected to be distributed uniformly, roughly following the dashed diagonal line (see Fig. 3). The figure indicates, however, that the p-values exhibit a sharp kink toward rejection of the null hypothesis at conventional levels.

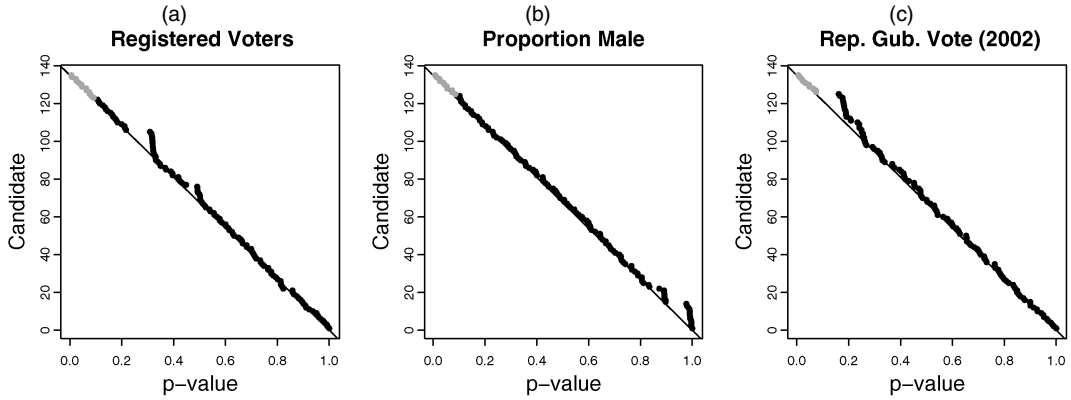

Figure 3. One-Tailed p-Values From Randomization Inference for Selected Pretreatment Covariates Not Affected by Page Placement. The p-values are arranged in order of their magnitude for each of the three pretreatment covariates: (a) number of registered voters, (b) proportion male in the total population, and (c) Republican vote share (Bill Simon) in the 2002 gubernatorial election. Gray dots denote the p-values which are $<.1$. Under the sharp null hypothesis of no page effects, the p-values are expected to be distributed uniformly, roughly following the diagonal line. The figure shows that unlike in Figure 2, this is indeed the case, indicating that these pretreatment covariates are not affected by the treatment.

4.2 Sensitivity Analyses

In this section we conduct two kinds of sensitivity analyses. First, we explore the possibility of relaxing the assumption of a constant additive treatment effect. So far, our estimation of confidence intervals assumed that the ballot page effect for a particular candidate is constant across districts. Although this assumption is shared by common parametric and nonparametric models (see Rosenbaum 2002b, p. 289), it may not be realistic in our application given the heterogeneity of California districts. Therefore, we estimate nonparametric confidence intervals for different subsets of the sample (see Rosenbaum 2002b, pp. 322–324, for a more general discussion). Significant differences in confidence intervals across those subsamples would suggest possible violation of the constant additive treatment effect assumption.

Figure 5 presents the estimated ballot page effects and their 90% nonparametric confidence intervals for major and selected minor candidates. For each candidate, the ballot page effect and their confidence intervals are estimated separately for Republican and Democratic districts. [Republican (Democratic) districts are defined as those districts where the proportion of registered Republican (Democratic) voters exceeds the proportion of Democratic (Republican) voters.] The results show that for major candidates, the effect sizes are similar across Republican and Democratic districts and large portions of the two estimated confidence intervals overlap with one another. For minor candidates, the story is similar, although there appears to be a noticeable difference between the two types of districts for Ramirez. With the exception of Ramirez, the constant additive treatment effect assumption appears to be plausible, at least with respect to these candidates and partisanship of the district.

We also attempted to test effects conditional on the total number of pages in a district, given the possibility that effects may not be similar when there are two pages compared with more than two pages. Unfortunately, identifying such heterogeneous effects was not possible, because in a number of districts where the total number of ballot pages exceeds two, candidates are not listed first for many permutations. We also investigated the heterogeneity of treatment effects of candidates by conducting rank tests on the estimated p-values by race and gender of candidates. Page effects do not appear to be related to candidate gender or race.

In our second sensitivity analysis, we examined the sensitivity of randomization inference to the choice of test statistics. In particular, we considered rank sum and median test statistics in addition to the difference-in-means and least squares test statistics. The rank sum test statistic is defined as $\sum_{i=1}^{121} T_i R(y_i)$, where $R(y_i)$ gives the rank of y_i among the 121 observed outcomes. The median test statistic represents the median value of the potential outcomes among the treated units. Figure 6 plots the logit-transformed p-values based on least squares, rank sum, and median test statistics against those based on the difference-in-means test statistic. The graphs show that similar inferences may be drawn from the different test statistics. The p-values based on these four test statistics are highly correlated. Regressing the logit-transformed p-values from the difference-in-means test statistic on those from other test statistics gives fitted lines that closely follow the 45-degree lines and yields a coefficient of determination approximately equal to .8. In this application, randomization inference does not appear particularly sensitive to the choice of test statistics.

4.3 Comparison With Conventional Estimators

Finally, we compare randomization inference with conventional estimators that do not fully incorporate the treatment assignment mechanism. Conventional analyses might assume complete or simple randomization of treatment assignment and compute mean differences in vote shares between when a candidate was and was not listed on the first page (Neyman 1923). Confidence intervals are often obtained by asymptotic normal approximation. This strategy may be reasonable in typical elections where the number of candidates is small relative to the number of districts (Ho and Imai 2004), but with 135 candidates running in the recall election, complete randomization is not ensured. Some candidates may be listed primarily on the first page in liberal Northern California while listed primarily on later pages of the ballot in conservative Southern California. As

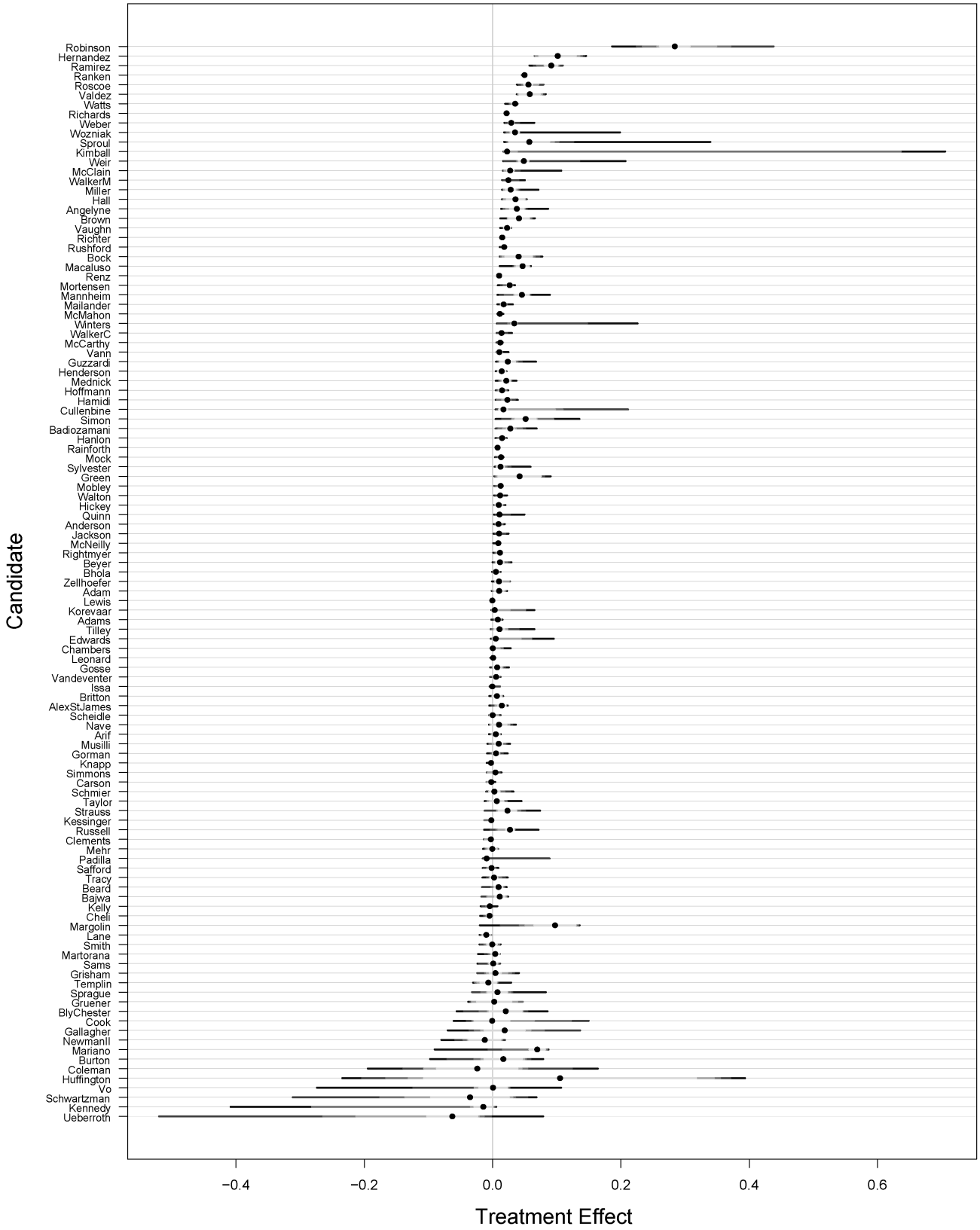

Figure 4. Nonparametric Confidence Intervals for Causal Effects of Being Listed on the First Ballot Page (— 60% interval; — 70% interval; — 80% interval; — 90% interval). For 17 candidates (Tsangares, Prady, Price, Pawlik, Palmieri, Pineda, Peters, Dole, Davis, Friedman, Forte, Foss, Fontanes, Farrell, Feinstein, Flynt, and Louie), the 90% confidence interval is not identifiable from the data. For 55 of the remaining 122 candidates, the 90% confidence interval exceeded the origin. Schwarzenegger, McClintock, Bustamante, Camejo, and Kunzman are excluded, to keep the scale for remaining candidates comparable.

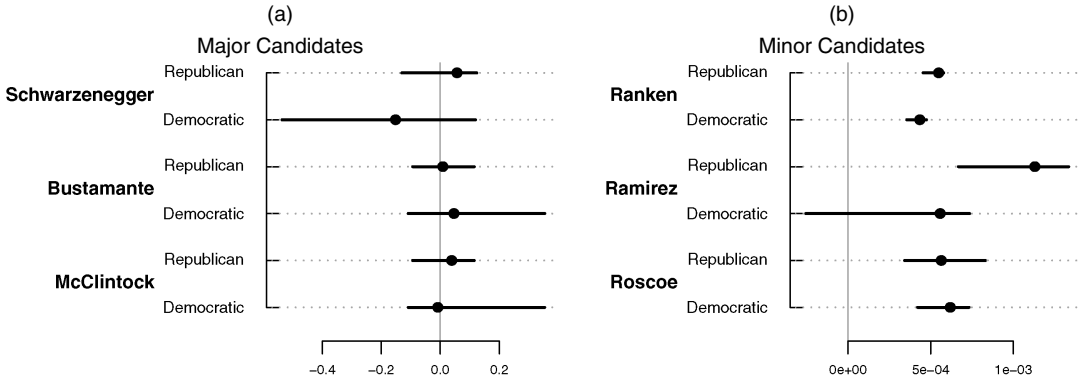

Figure 5. Estimated Ballot Page Effects for Major (a) and Selected Minor (b) Candidates for Republican and Democratic Districts. The graphs display the estimated ballot page effects (dark dots) and their 90% confidence intervals for major and selected minor candidates. The ballot page effects and their confidence intervals are estimated separately for Republican and Democratic districts. Republican (Democratic) districts are those where the proportion of registered Republican (Democratic) voters exceeds the proportion of Democratic (Republican) voters.

a result, estimates that assume complete randomization, ignoring the randomization-rotation procedure, may be confounded. Moreover, standard variance calculations that assume simple random assignment are invalid because of the systematic rotation of the California alphabet lottery. This means that unlike randomization inference, confidence intervals based on conventional estimators are likely to have incorrect coverage probabilities.

We investigate how randomization inference compares to two conventional estimators: linear least squares, which models vote shares, and the binomial generalized linear model (GLM) with a logit link and overdispersion, which models vote counts (McCullagh and Nelder 1989). We make the comparison twice, with and without covariates, to account for potential confounding effects due to incomplete randomization. The comparison between least squares and randomization inference is based on identical test statistics and the same covariate set defined in Section 2.2. The difference is that the reference distribution is derived from either the randomization of the treatment or the asymptotic normal approximation based on the linear model. We also note that these two conventional estimators may not be the best available parametric methods. Rather, we use them merely to compare the results of randomization inference with estimators frequently used by applied researchers, which are based on the same test statistics.

The upper part of Table 1 presents 90% confidence intervals from randomization inference and two conventional parametric analyses for the top three candidates. The results with and without covariance adjustment are given. Schwarzenegger was the Republican winner of the election and Bustamante was its runner-up and the main Democratic candidate. The results show appreciable differences between randomization inference and conventional estimators. For example, the estimates based on least squares regression without covariates imply that Schwarzenegger gained roughly 1–10 percentage points due to being on the first page. In contrast, the confidence interval based on randomization inference using the same test statistic is significantly wider and contains the origin. Even when controlling for the covariates, the confidence intervals based on the two methods differ significantly for Schwarzenegger.

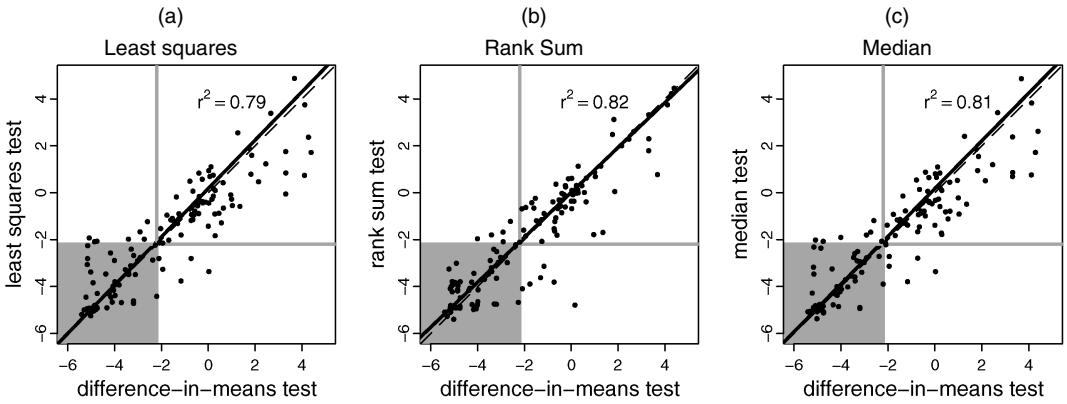

Figure 6. Insensitivity of Randomization Inference to the Choice of Test Statistics. Each graph plots the logit-transformed p-values based on three alternative test statistics against those based on the difference-in-means test statistic: (a) least squares; (b) rank sum; (c) median. The p-values are computed separately for each candidate using the alphabet randomization. The dashed lines show 45-degree lines, and the solid lines represent the least squares fit obtained by regressing the logit transformed p-values from the difference-in-means test statistic on those from an alternative test statistic. (r^2 represents the value of the coefficient of determination for the fitted least squares.) These two lines are closely aligned, indicating lack of sensitivity of randomization inference to the choice of test statistics in this application. Several candidates with p-values of 1 are omitted from the graphs. The candidates falling in the gray area have p-values $<.1$ for both test statistics.

Table 1. Comparison of Randomization Inference With Conventional Estimates That Do Not Incorporate the Treatment Assignment Mechanism

	Without covariates			*With covariates*		
	Least squares regression	*Binomial GLM with logit link*	*Randomization inference*	*Least squares regression*	*Binomial GLM with logit link*	*Randomization inference*
Major candidates						
Schwarzenegger	1.09 9.63	−1.23 7.53	−23.72 19.90	−2.97 .21	−4.98 −2.05	−6.44 6.87
Bustamante	−8.46 .54	−5.37 4.04	−20.07 20.31	−1.12 1.78	.96 3.01	−5.86 5.64
McClintock	.50 3.09	−1.10 1.24	−3.47 6.36	1.56 3.25	.29 2.05	.36 3.57
All candidates						
Positive effects	56 (41%)	63 (47%)	55 (41%)	50 (37%)	59 (44%)	47 (35%)
Negative effects	11 (8%)	8 (6%)	4 (3%)	8 (6%)	17 (13%)	2 (1%)
Null effects	68 (50%)	64 (47%)	59 (44%)	77 (57%)	59 (44%)	64 (47%)
Unidentified	0 (0%)	0 (0%)	17 (13%)	0 (0%)	0 (0%)	22 (16%)

NOTE: The left columns give the results without covariate information, and the right columns give the results using the covariates. The upper part of the table shows the 90% confidence interval for the causal effect of page placement on major candidates' vote shares. Randomization inference with covariates uses the test statistic defined in (9). When fitting the binomial GLM, we allow for overdispersion, and the results presented here are transformed to the scale of vote shares. The lower part of the table presents the summary of the results for all candidates and comparison of least squares and binomial GLM with corresponding randomization inference. The figures show the number and percentage of candidates whose 90% confidence intervals fall in each of the four categories: strictly positive ("Positive effects"), strictly negative ("Negative effects"), containing zero ("Null effects"), and not identified ("Unidentified").

We find similar differences between randomization inference and the binomial GLM. When controlling for the covariates, the result of the binomial GLM indicates that Schwarzenegger might have *lost* votes by 2–5 percentage points when listed on the first page. The results for Bustamante and McClintock differ less across three methods than do those for Schwarzenegger. Nevertheless, we observe some noticeable differences; for example, when controlling for the covariates, the binomial GLM shows a significant positive effect for Bustamante, whereas randomization inference and least squares regression do not.

The lower part of Table 1 summarizes the results for all 135 candidates. For example, randomization inference without (with) covariates detects significantly positive effects for 55 (47) candidates and significantly negative effects for 4 (2) candidates, whereas for 17 (22) candidates, confidence intervals are not identified. These results differ significantly from those based on the two conventional estimators. Regardless of whether one controls for covariates and despite the fact that the two methods use identical test statistics, substantive conclusions based on randomization inference agree with those based on least squares regression for only about 65% of the candidates. Moreover, conventional estimators detect significantly negative effects for a larger number of candidates than does randomization inference, contradicting the results of earlier studies. For example, the results based on the binomial GLM with covariates suggest that 17 candidates lost votes from being placed on the first page, whereas randomization inference with covariates indicate there are only two such candidates. Finally, the confidence interval based on randomization inference will be unidentified when the data are not informative (e.g., Imbens and Rosenbaum 2005). In our application, this is true for the candidates whose name appears on the first ballot page in only a handful of districts. In contrast, the conventional estimators identify significant effects for several of these candidates by making parametric assumptions.

In general, conventional parametric confidence intervals tend to be shorter than the nonparametric counterparts. Figure 7 compares the log length of the 90% parametric (linear least squares with and without covariates) and corresponding 90% nonparametric confidence intervals. We exclude those candidates for whom nonparametric confidence intervals are not identified. For comparison, we also conduct randomization inference with covariance adjustment from (9). Irrespective of whether or not the covariates are included, the length of the parametric confidence intervals tends to be substantially shorter than the nonparametric counterparts. The few dots below the 45 degree line represent candidates for whom the nonparametric confidence interval is shorter than the parametric counterpart. These candidates were listed on the first page in very few districts.

Finally, we compare the p-value curves of nonparametric and parametric estimators. The p-value curve can be obtained by plotting the null value of ballot page effect, τ_0, against its corresponding p-value. The curve is a step function because the total number of treatment assignment combinations is finite. As before, we use the alphabet letters as units of randomization, following the actual procedure of the recall election. For comparison, we also compute the p-value curve using the candidates as units of randomization, while maintaining the rotation procedure. The advantage of the latter approach is that it makes exact computation possible.

Figure 8 presents the p-value curves for Schwarzenegger with and without covariance adjustment. The p-value curves based on candidate and alphabet randomization (in solid gray and black lines, resp.) trail each other closely in both panels, suggesting that for Schwarzenegger little is lost by using candidates rather than alphabet letters as the unit of randomization as long as the rotation procedure is modeled. The figure also plots the p-value curve based on the asymptotic normal approximation of linear least squares regressions (with and without covariance adjustment). These parametric p-value curves deviate significantly from the nonparametric counterparts, suggesting that the large-sample normal approximation may not be appropriate in this application.

5. CONCLUSION

In this article we have illustrated how Fisher's exact test can be generalized to conduct (nonparametric) randomization inference for randomized natural experiments. For ethical and practical reasons, social scientists and policy makers can rarely conduct classical randomized experiments in real-world set-

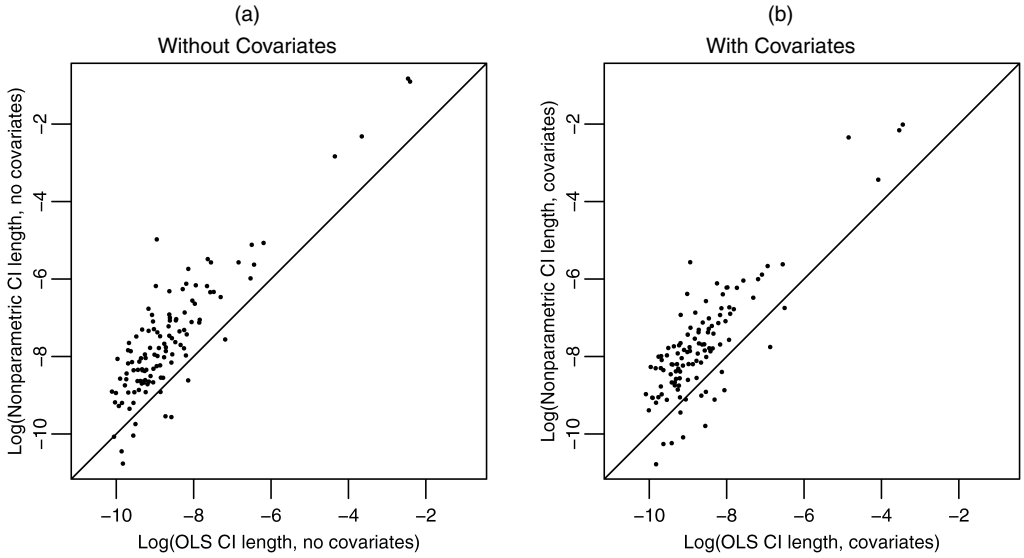

Figure 7. Comparing Length of 90% Confidence Intervals by Randomization Inference and Linear Least Squares. The log length of 90% nonparametric confidence intervals is on the vertical axis, whereas that of 90% parametric confidence intervals is on the horizontal axis. Plots are based on linear least squares (a) with and (b) without covariance adjustment. The plot omits 17 (22) candidates for which the nonparametric confidence interval without (with) covariates is not identified.

tings. Therefore, natural experiments such as the California alphabet lottery provide a rare opportunity for researchers and policy makers to draw causal inferences about particular quantities of interest while maintaining both internal and external validity. Randomization inference allows researchers to directly incorporate exact randomization procedures of natural experiments as the basis of statistical inference without introducing unnecessary distributional assumptions. We have shown that by inverting Fisher's exact test, accurate nonparametric confidence intervals can be obtained. In our analysis of the 2003 California recall election, we find that the results based on conventional estimators differ appreciably from those based on randomization inference that capitalizes on the systematic rotation of the treatment with a random start. Although the recall election was a peculiar election with an unprecedented number of candidates, our analysis demonstrates that when modeled appropriately, the results are consistent with those of other California general elections.

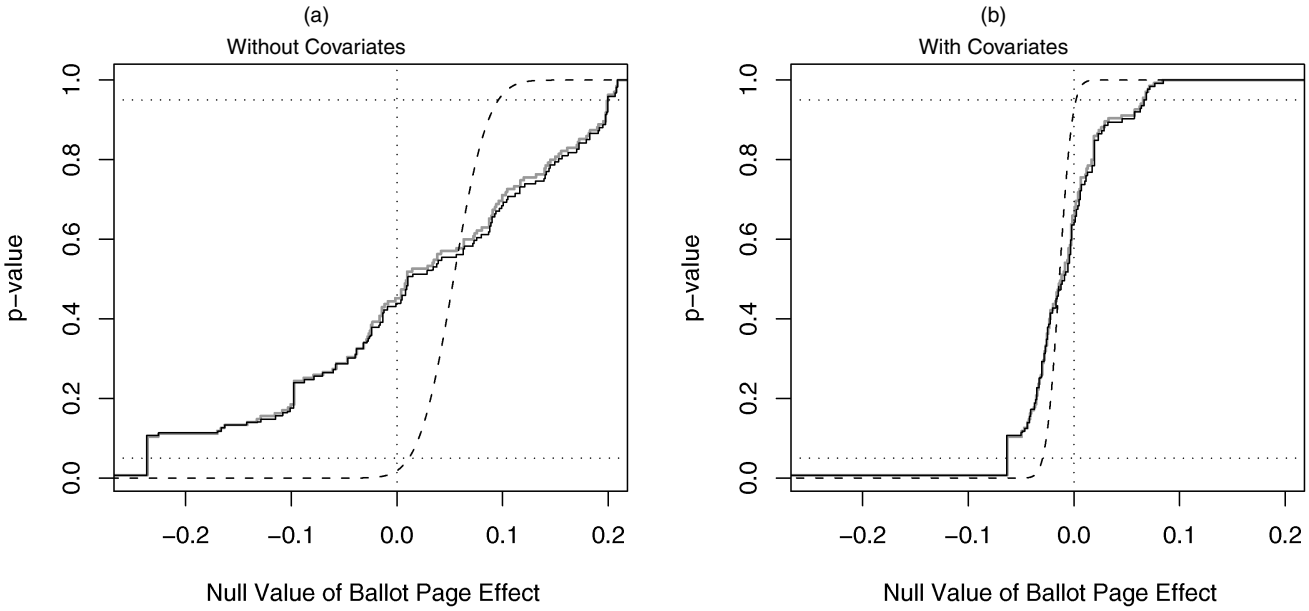

Figure 8. Comparison of p-Value Curves Between Parametric and Nonparametric Estimators. The figure plots (one-tailed) p-values against the corresponding null value of the ballot page effect, τ_0, for Arnold Schwarzenegger. Part (a) presents the p-value curve without covariance adjustment. Part (b) presents the p-value curve with covariance adjustment. The dashed lines represent the p-value curve from the linear least squares. For nonparametric inference, the black solid lines are based on alphabet randomization, whereas the gray solid lines are based on the candidate randomization. The vertical dotted line represents a treatment effect of 0, and the horizontal dotted lines represent bounds for a 90% confidence interval.

APPENDIX: COMPUTATION OF NONPARAMETRIC CONFIDENCE INTERVALS

Computing randomization-based nonparametric confidence intervals is approximately equivalent to finding the roots for the following nonlinear equations:

$$\begin{aligned} f(\tau_{\mathrm{U}}) &= \Pr\bigl(W_{\tau_{\mathrm{U}}}(\mathbf{T}) \geq W(\mathbf{t})\bigr) - \frac{\alpha}{2}, \\ g(\tau_{\mathrm{L}}) &= \Pr\bigl(W_{\tau_{\mathrm{L}}}(\mathbf{T}) \geq W(\mathbf{t})\bigr) - 1 + \frac{\alpha}{2}, \end{aligned} \tag{A.1}$$

where τ_{U} and τ_{L} are upper and lower bounds of the confidence set. To solve these nonlinear equations, we use a simple extension of the bisection algorithm, which in our current application is more appropriate than other nonlinear optimization techniques, such as Newton–Raphson algorithms. This is because the latter methods require that the objective functions be continuous and/or differentiable, whereas $f(\cdot)$ and $g(\cdot)$ are discrete in our application. Our extension here is that we use Monte Carlo simulation to approximate the values of the functions, $f(\cdot)$ and $g(\cdot)$, the exact evaluations of which are computationally demanding.

[Received September 2004. Revised April 2005.]

REFERENCES

Aldrich, J. H. (1993), "Rational Choice and Turnout," *American Journal of Political Science*, 37, 246–278.

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," *American Economic Review*, 80, 313–336.

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Bain, H. M., and Hecock, D. S. (1957), *Ballot Position and Voter's Choice*, Detroit: Wayne State University Press.

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

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

Cox, D. R., and Hinkley, D. (1979), *Theoretical Statistics*, London: Chapman & Hall/CRC.

Cox, D. R., and Reid, N. (2000), *The Theory of the Design of Experiments*, New York: Chapman & Hall.

Darcy, R. (1986), "Position Effects With Party Column Ballots," *Western Political Quarterly*, 39, 648–662.

Darcy, R., and McAllister, I. (1990), "Ballot Position Effects," *Electoral Studies*, 9, 5–17.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Forsythe, R., Myerson, R. B., Rietz, T. A., and Weber, R. J. (1993), "An Experiment on Coordination in Multi-Candidate Elections: The Importance of Polls and Election Histories," *Social Choice and Welfare*, 10, 223–247.

Garrett, E. (2004), "Democracy in the Wake of the California Recall," *University of Pennsylvania Law Review*, 152, 239–284.

Gelman, A., King, G., and Boscardin, J. W. (1998), "Estimating the Probability of Events That Have Never Occurred: When Is Your Vote Decisive," *Journal of the American Statistical Association*, 93, 1–9.

Gold, D. (1952), "A Note on the 'Rationality' of Anthropologists in Voting for Officers," *American Sociological Review*, 17, 99–101.

Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004), "Randomization Inference With Imperfect Compliance in the Ace-Inhibitor After Anthracycline Randomized Trial," *Journal of the American Statistical Association*, 99, 7–15.

Hansen, B. E. (2003), "Recounts From Undervotes: Evidence From the 2000 Presidential Election," *Journal of the American Statistical Association*, 98, 292–298.

Ho, D. E., and Imai, K. (2004), "Estimating Causal Effects of Ballot Order From a Randomized Natural Experiment: California Alphabet Lottery, 1978–2002," unpublished manuscript, available at *http://imai.princeton.edu/research/alphabet.html*.

Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–960.

Imai, K., and King, G. (2004), "Did Illegal Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?" *Perspectives on Politics*, 2, 537–549.

Imai, K., and van Dyk, D. A. (2004), "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866.

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Ser. A, 168, 109–126.

Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: Wiley.

Kershaw, S. (2003), "Recall Voters Face an Intricate Ballot, and, Indeed, Chads," *The New York Times*, October 6, 2003, A1.

Krosnick, J. A., Miller, J. M., and Tichy, M. P. (2003), "An Unrecognized Need for Ballot Reform," in *Rethinking the Vote*, eds. A. W. Crigler, M. R. Just, and E. J. McCaffery, New York: Oxford University Press, pp. 51–74.

Lange, K. (1999), *Numerical Analysis for Statisticians*, New York: Springer-Verlag.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

Miller, J. M., and Krosnick, J. A. (1998), "The Impact of Candidate Name Order on Election Outcomes," *Public Opinion Quarterly*, 62, 291–330.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," *Statistical Science*, 5, 465–480.

Pocock, S. J., and Simon, R. (1975), "Sequential Treatment Assignment With Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–115.

Riker, W. H., and Ordeshook, P. C. (1968), "A Theory of the Calculus of Voting," *American Political Science Review*, 62, 25–42.

Rosenbaum, P. R. (1996), Comments on "Identification of Causal Effects Using Instrumental Variables," by J. D. Angrist, G. W. Imbens, and D. B. Rubin, *Journal of the American Statistical Association*, 91, 465–468.

——— (2002a), "Attributing Effects to Treatment in Matched Observational Studies," *Journal of the American Statistical Association*, 97, 183–192.

——— (2002b), "Covariance Adjustment in Randomized Experiments and Observational Studies" (with discussion), *Statistical Science*, 17, 286–327.

——— (2002c), *Observational Studies*, New York: Springer-Verlag.

Rubin, D. B. (1990), Comments on "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," by J. Splawa-Neyman (translated from the Polish and edited by D. M. Dabrowska and T. P. Speed), *Statistical Science*, 5, 472–480.

Scott, J. W. (1972), "California Ballot Position Statutes: An Unconstitutional Advantage for Incumbents," *Southern California Law Review*, 45, 365–395.

Shaughnessy, J. J., Zechmeister, E. B., and Zechmeister, J. S. (2002), *Research Methods in Psychology*, New York: McGraw-Hill.

Starr, N. (1997), "Nonrandom Risk: The 1970 Draft Lottery," *Journal of Statistics Education*, 5, *http://www.amstat.org/publications/jse/v5n2/datasets.starr.html*.

Tomz, M., and Van Houweling, R. P. (2003), "How Does Voting Equipment Affect the Racial Gap in Voided Ballots?" *American Journal of Political Science*, 47, 46–60.

U.S. General Accounting Office (2001), "Statistical Analysis of Factors That Affected Uncounted Votes in the 2000 Presidential Election," Report to the Ranking Minority Member, Committee on Government Reform, House of Representatives.

Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Jr., Herron, M. C., and Brady, H. (2001), "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida," *American Political Science Review*, 95, 793–810.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1, 80–83.

Wolter, K. M. (1984), "An Investigation of Some Estimators of Variance for Systematic Sampling," *Journal of the American Statistical Association*, 79, 781–790.