Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments*

Avi Feller and Chris C. Holmes†

This Version: July 20, 2009
Original Version: June 23, 2009

Abstract

Randomized experiments have become increasingly important for political scientists and campaign professionals. With few exceptions, these experiments have addressed the overall causal effect of an intervention across the entire population, known as the average treatment effect (ATE). A much broader set of questions can often be addressed by allowing for heterogeneous treatment effects. We discuss methods for estimating such effects developed in other disciplines and introduce key concepts, especially the conditional average treatment effect (CATE), to the analysis of randomized experiments in political science. We expand on this literature by proposing an application of generalized additive models (GAMs) to estimate non-linear heterogeneous treatment effects. We demonstrate the practical importance of these techniques by re-analyzing a major experimental study on voter mobilization and social pressure and a recent randomized experiment on voter registration and text messaging from the 2008 US election.

1 Introduction

In the last decade, political scientists and campaign professionals have conducted well over one hundred randomized experiments, often involving tens of thousands of participants or more [Druckman et al., 2006, Green and Gerber, 2008]. In these experiments, the main quantity of interest is the *average treatment effect* (ATE), a measure of the overall impact of a treatment. Armed with the ATE, political scientists have addressed questions like "How much does a Get Out The Vote (GOTV) campaign increase voter turnout?" or "Is a phone call or an in-person conversation more effective at increasing candidate support?" [Gerber and Green, 2008]. Campaign professionals often call these basic results the "toplines" of an experiment.

Increasingly, however, researchers and campaigns are interested in more complex questions like "Does a GOTV phone call increase turnout more for men or women?" or "How does the effectiveness of a GOTV phone call change at different ages?" While statistical methods for analyzing such *heterogeneous*

*We would like to thank Jim Alt, Ian Crawford, Ray Duch, Andrew Gelman, Don Green, Guido Imbens, Chris Mann, David Nickerson, Simon Quinn, Todd Rogers, Aaron Strauss, and seminar participants at the University of Oxford for their helpful comments. We would especially like to thank Chris Kennedy, Rock the Vote, and Catalist for sharing experimental data. AF would like to thank the Rhodes Trust for financial support. CCH would like to thank the Oxford Man Institute for supporting this research.

†Department of Statistics, University of Oxford, Oxford, OX1 3TG, and Oxford Man Institute, University of Oxford, Oxford, OX1 4EH; email: avi.feller@gmail.com

treatment effects are widely employed in other disciplines [Rothwell, 2005, Imbens and Wooldridge, 2009], they remain uncommon in the field experiments literature and are little used elsewhere in political science.[1] Those researchers and campaign professionals who do investigate treatment effect heterogeneity generally rely on *ad hoc* analyses, often by looking at simple differences in treatment effects across the demographic breakdown of an experiment. Extensive reviews in medical statistics, however, suggest that without appropriate methods for analyzing these results, researchers run the risk of drawing erroneous conclusions about their experiments [Assmann et al., 2000, Pocock et al., 2002].

To resolve this problem, we apply methods developed in econometrics and medical statistics to randomized experiments in political science. Specifically, we describe the idea of the *conditional average treatment effect* (CATE), defined as the average treatment effect conditional on a given covariate or set of covariates, which facilitates the analysis of heterogeneous treatment effects [Imbens and Wooldridge, 2009]. We outline the use of standard regression techniques to extend the basic toplines idea to allow for a more detailed and careful study of experimental results.

Estimating conditional average treatment effects using regression, however, requires the specification of a functional form for treatment effect heterogeneity. As Beck and Jackman [1998] point out, "few social scientific theories offer any guidance as to functional form whatsoever." Following Beck and Jackman, we therefore utilize *generalized additive models* (GAMs), which are an extension of the standard regression model to allow for non-linear effects [Hastie and Tibshirani, 1990, Wood, 2006]. Unlike other common non- and semi-parametric approaches, GAMs can estimate both linear and non-linear effects simultaneously, yielding a powerful and flexible framework for the analysis of non-linear treatment effect heterogeneity with continuous covariates. This framework enables researchers to address questions like "How does the effectiveness of a GOTV phone call change at different ages?" without imposing a parametric form on the effect.

There is an extensive literature on treatment effect heterogeneity, especially in econometrics and biostatistics. The concept was first introduced in early statistical work as *treatment-covariate interaction*,[2] although most of this research focused on reducing heterogeneity rather than estimating it [Fisher, 1935, Johnson and Neyman, 1936, Johnson and Fay, 1950, Potthoff, 1964, Cox, 1958, 1984]. Educational researchers and biostatisicians soon expanded the statistical foundations to measure these differences in treatment effects across subjects [Cronbach and Snow, 1977], known in the medical literature as *subset analysis* [Byar, 1985, Dixon and Simon, 1991]. Econometricians followed with the structural models of Heckman [2001, 2008] [see also Björklund and Moffitt, 1987] and the econometric framework for program evaluation [Imbens and Wooldridge, 2009]. Although treatment effect heterogeneity has been a

[1] Notable exceptions using observational studies are Gelman and King [1994] and Gelman and Huang [2008].
[2] The terms *concommitant variables* and *intrinsic factors* were often used in place of covariate.

major research area in each of these fields for decades, the literatures have remained largely distinct, often reinventing advances in other disciplines.

The corresponding research in political science has largely focused on experiments in voter mobilization and political participation. Using a Bayesian framework, Horiuchi et al. [2007] walk through a detailed analysis of a randomized experiment with noncompliance and nonresponse. The authors highlight the importance of heterogeneity, but focus on differential effects across a continuous treatment variable rather than on differential effects across covariates. While Horiuchi et al. briefly address differences in treatment effects across subgroups of the population, they do not offer a structured approach for determining whether these differences are statistically significant. Another analysis of heterogeneity is recent work by Imai and Strauss [2009], who consider differences in treatment effects across covariates but tailor their approach to optimize voter mobilization campaigns. Like other work on optimal treatment assignment [Aitchison, 1970, Byar and Corle, 1977, Dehejia, 2005], their purpose is maximize the efficiency of the overall campaign rather than to estimate differences between types of voters. As such, their algorithmic approach is not designed to assess whether two subgroups are statistically distinct. Finally, Arceneaux and Nickerson [2009] analyze eleven randomized experiments to examine the heterogeneous effects of voter mobilization by underlying vote propensity. Their meta-analysis approach, however, does not employ a statistical framework for treatment effect heterogeneity and cannot be readily extended to other covariates of interest.

We begin our paper with a discussion of the statistical framework underlying the analysis of randomized experiments, known as the Neyman-Rubin Model, and introduce the idea of conditional average treatment effects in this context. We then give a brief introduction to generalized additive models and extend the idea of treatment effect heterogeneity to include non-linear interactions. We demonstrate the utility of allowing for treatment effect heterogeneity by analyzing two randomized field experiments. First, we re-analyze an experiment conducted by Gerber et al. [2008] in the 2006 Michigan primary election, which examines the effects of social pressure on voter turnout. Although the original analysis found no differences in treatment across subgroups, we re-analyze the same data using the idea of CATE and find strong evidence of treatment effect heterogeneity by vote propensity. Second, we analyze an experiment from the 2008 election cycle conducted by Rock the Vote, which uses text messages to remind young people to register to vote. We use this experiment to illustrate the interpretation of GAMs and the estimation of non-linear treatment effect heterogeneity. We conclude with a discussion of treatment effect heterogeneity and possibilities for future work.

2 Statistical Framework

2.1 Average Treatment Effect

Randomized experiments are powerful tools for assessing the causal effect of an intervention. In a randomized experiment, the experimental units (e.g. individuals or households) are randomly assigned to treatment or control groups. Since the treatment and control groups are identical except for the randomized assignment, we can attribute differences in the outcome (e.g. voting) to the causal effect of the treatment. The most common method for analyzing such experiments is the statistical framework known as the Rubin Causal Model or the Neyman-Rubin Model [Holland, 1986, Angrist et al., 1996], due to foundational work by Neyman [1923] and Fisher [1935] and important extensions by Cox [1958] and Rubin [1974, 1978].[3]

To illustrate the Neyman-Rubin Model, we first consider the causal effect of a randomly assigned binary treatment D, with $D_i = 1$ indicating random assignment to the treatment group and $D_i = 0$ indicating random assignment to the control group (the more general case of multiple treatments is described below). The central idea of the Neyman-Rubin model is that of a *potential outcome*, defined as $Y_i(1)$ if person[4] i is assigned to the treatment group, $D_i = 1$, and $Y_i(0)$ if person i assigned to the control group, $D_i = 0$. The treatment effect for person i is defined as the difference between the potential outcomes,

$$\tau_i = Y_i(1) - Y_i(0)$$

The problem, however, is that each person can only be assigned to either the treatment or control group, but not both, leading to what Holland [1986] calls the *fundamental problem of causal inference*. Randomization mitigates this problem because the treatment and control groups are constructed such that the groups are identical across observed and unobserved covariates. We can therefore compare averages across all individuals in the treatment and control groups, yielding the *average treatment effect* (ATE),

$$\text{ATE} = \mathbb{E}(\tau_i) = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0))$$

In other words, the difference between the means of the randomly assigned treatment and control groups yields an unbiased estimate of the average treatment effect even though the individual treatment effects, τ_i, can vary dramatically. Nominal confidence intervals for the estimate of ATE, $\hat{\tau}$, can be calculated under weak restrictions from the variance, $\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$, where S_D^2 is the sample variance and n_D is the

[3] Pearl [2000] offers a "competing" framework for causal inference based on graphical representations. Examining treatment effect heterogeneity in Pearl's framework remains an active area of research and is not addressed here.

[4] For ease of reading, we use "person" or "individual" in place of the more general "experimental unit." The logic holds for any experimental unit, such as a household or precinct.

number of individuals for group D [Hahn, 1998, Imbens and Wooldridge, 2009]. Additionally, a range of non-parametric tests, based largely on Fisher's exact test, can be used to evaluate the null hypothesis that $\hat{\tau} = 0$ [e.g. Rosenbaum, 2002, Hansen and Bowers, 2009].

In the case of a randomized experiment, the ATE estimate requires only mild assumptions, namely the exclusion restriction and the Stable Unit Treatment Value Assumption (SUTVA) [see Gerber and Green, 2008].[5] While these assumptions generally hold, a more common problem is that of noncompliance, where, for example, an individual was assigned to the treatment group but was not actually treated. Noncompliance and the related problem of nonresponse can lead to heavily biased ATE estimates [Horiuchi et al., 2007]. A large literature addresses these problems using estimators like the average treatment effect on the treated (ATT) and the local average treatment effect (LATE) [Imbens and Angrist, 1994, Gerber and Green, 2008]. These estimators are not explored here, but we note that in such a situation, the ATE described above is the Intention-to-Treat Effect (ITT) [Imai, 2005, Freedman, 2008b].

2.2 Conditional Average Treatment Effect

The ATE estimate can be thought of as an imprecise estimator of the individual treatment effect, τ_i. In sufficiently large experiments, it is possible to improve this precision by estimating treatment effects for a subpopulation that more closely resembles the individual treatment effect. We might believe, for example, that men and women respond very differently to the same treatment, where men show an eight percent increase in turnout but women only show a four percent increase. Although the ATE consistently estimates an overall effect of six percent, no individual values of τ_i are close to this estimate. We can instead explicitly consider differences in treatment effectiveness between men and women by estimating the *conditional average treatment effect* (CATE), which is the average treatment effect conditional on a given covariate or set of covariates [Imbens and Wooldridge, 2009]. The resulting estimator is

$$\begin{aligned} \text{CATE} = \mathbb{E}(\tau(x)) &= \frac{1}{n}\sum_{i=1}^{N} \mathbb{E}(Y_i(1) - Y_i(0)|X = x_i) \\ &= \mathbb{E}(Y_i(1)|X = x_i) - \mathbb{E}(Y_i(0)|X = x_i) \end{aligned}$$

where $\tau(x)$ is the conditional average treatment effect for all individuals with $X = x$. CATE for men is therefore $\mathbb{E}(\tau(\text{male})) = \mathbb{E}(Y_i(1)|\text{male}) - \mathbb{E}(Y_i(0)|\text{male})$, which is the difference between the conditional means for the treatment and control groups among men. Analogously, CATE for women is the difference between the conditional means for the treatment and control groups among women.

It is important to note that causal claims of CATE estimates in the Neyman-Rubin framework can

[5] The program evaluation literature frames the critical assumption as *unconfoundedness*, that the treatment assignment is conditionally independent of the outcome given the observed covariates. For a detailed discusion, see Imbens and Wooldridge [2009] and Rosenbaum and Rubin [1983].

only be made about treatment effects conditioned on variables that are measured prior to treatment or are intrinsic to the individual.[6] Consider a randomized experiment testing a partisan door-to-door voter mobilization drive, which also includes a pre-treatment survey in which respondents are asked their party identification. In this case, it would be reasonable to explore differences in treatment effectiveness by reported partisanship. If the survey had been conducted after the treatment, however, reported partisanship could be influenced by the treatment itself. In this case, reported partisanship would become another outcome variable, and the causal interpretation is unclear for estimating the effect of an intervention for one outcome conditional on a separate outcome. Note that this reasoning does not apply to intrinsic variables like gender and age.

The sample estimates for CATE require no additional assumptions over ATE and, like ATE, are unbiased estimates of the average treatment effect conditional on a set of covariates. In this way, it is possible to estimate the conditional average treatement effect for any combination of covariates, which is akin to the popular practice of comparing "cross-tabs" of experimental results. There is a danger here, however. While a researcher might estimate the CATE for all women, say, estimating the CATE for left-handed, unmarried mothers without a strong reason *ex ante* would likely overfit the data. Overfitting has proved to be a major problem in medical studies; Assmann et al. [2000] reviewed fifty major articles and found that most of the surveyed papers report differences in treatment effects across covariates despite sufficient statistical power to detect such differences and without appropriate statistical methods [Pocock et al., 2002]. Statistical tests are therefore important for determining whether treatment effects significantly differ across a set of covariates. One approach proposed by Crump et al. [2008] is to use a non-parametric test for treatment effect heterogeneity, requiring only mild assumptions [see also Crump et al., forthcoming]. Although this and other model-free tests are important, they are limited in application. At the cost of stronger assumptions, we therefore evaluate treatment effect heterogeneity using the more flexible regression framework.

2.3 Regression Framework

Regression analysis has become the standard method for analyzing randomized experiments. Under appropriate assumptions, regression yields consistent estimators for ATE and CATE, which are also unbiased in certain situations [Freedman, 2008b]. Furthermore, including covariates in the regression can also reduce the residual variance, in turn reducing the standard errors of the treatment effect estimates [Angrist and Pischke, 2009]. The use of standard regression techniques, however, requires much stronger assumptions than the straightforward sample estimates, namely additivity, homoskedasticity,

[6] For adjustments on post-treatment variables, see the literature on post-stratification, especially Little [1993] and Frangakis and Rubin [2002].

and linearity. Freedman [2008b] shows that the Neyman-Rubin Model does not generally justify such assumptions and that the resulting estimates can yield biased results for even moderately large experiments. Nonetheless, the bias associated with covariate-adjusted multiple regression treatment effect estimates is on the order of $O(1/n)$, tending to zero in sufficiently large experiments. The same reasoning does not hold for logit and probit models, however, which are especially prone to bias if the added assumption of an underlying logistic or probit function does not hold [Freedman, 2008a]. The inconsistency of logit and probit models is unfortunate since many political science experiments have a binary response. We therefore rely on the so-called Linear Probability Model (LPM), which is standard Ordinary Least Squares (OLS) regression with a binary outcome. This approach has significant drawbacks, especially the possibility of predicting response probabilities less than zero or greater than one. Nonetheless, LPM is more robust than limited dependent variable methods for randomized experiments if the treatment effect is small and the probability of success is not extreme [Angrist, 2001, Angrist and Pischke, 2009]. Regardless of regression method, however, it is important to compare regression estimates against sample calculations whenever possible.

With these caveats, the standard regression representation of the Neyman-Rubin Model is

$$Y_i = \alpha + \tau D_i + \sum_j \beta_j x_{i,j} + \epsilon_i \tag{1}$$

where Y_i is the observed outcome for person i, τ is the ATE, D_i is a dummy variable for treatment effect, $x_{i,j}$ are the covariates, and ϵ_i is the error term, which is Normal with a mean of zero and an unknown variance σ^2. The estimate for ATE is the coefficient for D_i, $\hat{\tau}$, which we can test for statistical significance using a standard t-test. If the experiment is properly balanced on covariates, this approach yields an identical estimate to the standard topline results, but with the additional benefit of regression standard errors.

We estimate the conditional average treatment effects in a simple extension to regression (1) by introducing an interaction term between the dichotomous covariate of interest and the main treatment effect, D [e.g. Gelman, 2009]. This approach is known as *subgroup analysis* in the medical statistics literature [Byar, 1985, Dixon and Simon, 1991]. Returning to the estimation of CATE for men and women, the regression becomes

$$\begin{aligned} Y_i &= \alpha + \tau D_i + \sum_j \beta_j x_{i,j} + \gamma\, \texttt{male}_i + \delta\, \texttt{male}_i D_i + \epsilon_i \\ &= \alpha + (\tau + \delta\, \texttt{male}_i) D_i + \sum_j \beta_j x_{i,j} + \gamma\, \texttt{male}_i + \epsilon_i \end{aligned}$$

Therefore, the estimate of CATE for men is $\widehat{\tau(\texttt{male})} = \hat{\tau} + \hat{\delta}\,\texttt{male}_i$, and for women is $\tau(\widehat{\texttt{female}}) = \hat{\tau}$. The test for statistically significant heterogeneity in treatment effects across the two groups is simply a t-test for the significance of the coefficient $\hat{\delta}$. This calculation is readily extended to a variable with multiple categories, such as race, using an F-test for the joint significance of the interaction terms.

An alternative specification estimates separate effects for each subgroup, again $\widehat{\tau(\texttt{male})}$ and $\tau(\widehat{\texttt{female}})$, without estimating the main treatment effect:

$$Y_i = \alpha + \texttt{male}_i \cdot \tau(\texttt{male}) \cdot D_i + \texttt{female}_i \cdot \tau(\texttt{female}) \cdot D_i + \sum_j \beta_j x_{i,j} + \epsilon_i \tag{2}$$

where $\texttt{male}_i$ and $\texttt{female}_i$ are dummy variables for gender. In this specification, the statistical significance of the interaction is tested using an F-test on the joint significance of $\widehat{\tau(\texttt{male})}$ and $\tau(\widehat{\texttt{female}})$.

One significant advantage of the regression framework is the ability to estimate continuous heterogeneous treatment effects, which is not possible using a standard cross-tabs approach. In this way, the interaction term is analogous to the idea of the *marginal treatment effect* [Heckman, 2001, Heckman and Vytlacil, 2005]. If we assume a linear conditional average treatment effect for a continuous covariate like age, the resulting regression is again of the form

$$Y_i = \alpha + (\tau + \delta\,\texttt{age}_i)D_i + \sum_j \beta_j x_{i,j} + \gamma\,\texttt{age}_i + \epsilon_i \tag{3}$$

The estimate of CATE for age is $\widehat{\tau(\texttt{age})} = \hat{\tau} + \hat{\delta}\texttt{age}_i$. Again, the significance of the linear interaction is tested using a standard t-test on $\hat{\delta}$. The interpretation of the continuous conditional average treatment effect is the same as for any other continuous-binary interaction term, although care should be taken not to extrapolate beyond the support of the covariate.

As with typical regression practice, higher-order terms can be used to approximate non-linearities for both the main and interaction effects. Such higher-order polynomials are especially susceptible to outliers and have several other undesirable statistical properties [Silverman, 1985]. A more general framework allows for non-linear effects without the need to specify a given functional form. The resulting regression is

$$\begin{aligned} Y_i &= \alpha + \tau D_i + \sum_j \beta_j x_{i,j} + f(\texttt{age}_i) + g(\texttt{age}_i)D_i + \epsilon_i \\ &= \alpha + (\tau + g(\texttt{age}_i))D_i + \sum_j \beta_j x_{i,j} + f(\texttt{age}_i) + \epsilon_i \end{aligned}$$

where $f(\texttt{age}_i)$ and $g(\texttt{age}_i)$ are "smooth" non-linear functions. Simultaneously estimating multiple non-

linear functions is beyond the scope of the ordinary regresion framework. Fortunately, a well-developed statistical method known as generalized additive models (GAMs) is designed for this task.

2.4 Generalized Additive Models

There is an extensive literature on generalized additive models and their application in wide range of fields [for surveys, see Hastie and Tibshirani, 1990, Ruppert et al., 2003, Wood, 2006]. We describe the method only briefly and direct interested readers to Beck and Jackman [1998], who give a useful introduction to GAMs for political scientists. While many different statistical packages estimate GAMs, we use the `mgcv` package for `R` [Wood, 2006].[7]

To understand GAMs, first consider the standard linear model,

$$Y_i = \alpha + \sum_j \beta_j x_{i,j} + \epsilon_i$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$. In additive models, we instead have

$$Y_i = \alpha + \sum_j m_j(x_{i,j}) + \epsilon_i$$

where $m_j(\cdot)$ is a "smooth" non-linear function and ϵ_i again has $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$. We can generalize these models, as for generalized linear models [McCullagh and Nelder, 1989], so that the response relates to the predictors via a *link function*, such as the logit function. When there are multiple smooth terms, GAMs estimate all smooth functions jointly by an iterative process known as *backfitting*, and produce the non-linear functions $\hat{m}_j(x_{i,j})$, which have zero mean and "coefficient" one by construction. Each smooth function, typically a spline, is then displayed graphically and is easily interpretable, as demonstrated in the example below.

Well-developed theory provides tests on the significance of each smooth term and the overall goodness-of-fit. One important quantity is the *effective degrees of freedom* (EDF), which is a measure of the degrees of freedom used by the smoothing term. If a smoothing term has EDF = 2, for example, than it would consume the same number of degrees of freedom in a model as a quadratic effect (with appropriate marginal terms). The degree of smoothing in `mgcv` is computed automatically, although the level of complexity as defined by the degrees of freedom can be adjusted to fit the data, such that more complex additive effects warrant higher degrees of freedom. A second important quantity is the F-value of a smoothing term. Splines and other so-called *linear smoothers* can be represented by a smoothing matrix,

[7] Ruppert et al. [2003] introduced the `R` package `semipar` specifically for semiparametric regression, which is a useful alternative to the `mgcv` package in many situations.

$\mathbf{S}$, such that $\hat{m}(\mathbf{x}) = \mathbf{Sy}$, in which each row of $\mathbf{S}$ gives the smoothing weights for each data point. The F-value therefore corresponds to the joint significance of the columns of $\mathbf{S}$, enabling straightforward significance tests for each smooth term [Hastie and Tibshirani, 1990]. It is similarly possible to use ANOVA to test whether a non-linear term should be used in place of a linear term.[8] Finally, GAMs can have both linear and non-linear effects in the same model, reducing to a standard regression framework if the data only support linear terms. A model in which there are both linear and non-linear effects is known as a semiparametric model [Ruppert et al., 2003].

The basic semiparametric model for the Neyman-Rubin Model is an extension of the causal regression in (1) to allow for non-linear covariate effects. First, we divide the covariates into two groups, x_j and z_k. We treat x_j like standard linear regression coefficients but allow z_k to vary non-linearly. The resulting semiparametric model is

$$Y_i = \alpha + \tau D_i + \sum_j \beta_j x_{i,j} + \sum_k m_k(z_{i,k}) + \epsilon_i$$

where $m_k(\cdot)$ are non-linear functions approximated by splines. We can further extend this model to allow for an interaction term, which is similar to research on non- and semi-parametric analysis of covariance [Eubank et al., 1995, Hunsberger and Follmann, 2001]. The resulting semiparametric model is

$$Y_i = \alpha + (\tau + m_2(\texttt{age}_i))D_i + \sum_j \beta_j x_{i,j} + m_1(\texttt{age}_i) + \epsilon_i \tag{4}$$

where $m_1(\cdot)$ and $m_2(\cdot)$ are non-linear functions.[9] The non-linear estimate of CATE for age is $\hat{\tau}+\hat{m_2}(\text{age}_i)$, which can be displayed graphically. To test whether or not there is statistically significant treatment effect heterogeneity by age, we simply test the significance of the F-value of the smooth term, $\hat{m_2}(\texttt{age}_i)$, which is part of the standard output. To test whether $\hat{m_2}(\texttt{age}_i)$ should be replaced by a parametric term, we estimate a separate regression substituting the parametric term for the smooth term, and use ANOVA to test differences between the nested models. We note that we can also estimate treatment effect heterogeneity for multiple continuous variables simultaneously, such as CATE for age and CATE for income. Additionally, multivariate smoothers can calculate conditional effects for two continuous variables at once, such as CATE for the interaction of income and age, resulting in a three-dimensional treatment effect surface. While such multivariate smoothers are only recommended with large data sets, the theory and intrepretation of GAMs remain relatively unchanged even in high-dimensional space [Wood, 2006].

The literature on GAMs and their use in modeling interactions is broad and well-developed. While

[8] Hastie and Tibshirani [1990] develop an automated approach, known as BRUTO, to identify when non-linear terms significantly improve model fit. This procedure is most useful for a large number of covariates, which is a situation not addressed here.

[9] Example `R` code for this equation using the `mgcv` library is `gam(y ~ treatment + s(x) + s(x, by = treatment))`

GAMs are not universally applicable [for a discussion, see Beck and Jackman, 1998], they can prove to be powerful modeling tools when used appropriately.

3 Social Pressure and Voter Turnout Experiment

3.1 Original Experiment

In order to demonstrate the importance of treatment effect heterogeneity, we examine a major study published in the *American Political Science Review*. Gerber et al. [2008] conducted a massive direct mail randomized experiment with 344,084 voters in 180,002 households across Michigan during the August 2006 primary election. Among other selection criteria, all 344,084 people had voted in the November 2004 general election. Twenty thousand households were randomly assigned to each of four treatment groups and the remaining 100,000 households were assigned to the control group. Each household in the treatment group received one of four mailings, which are shown in the Appendix.

- The *Civic Duty* mailing solely emphasizes civic duty and is meant as a baseline for comparison with the other treatments.
- The *Hawthorne* mailing states "YOU ARE BEING STUDIED!" and informes voters "that their voting behavior would be examined by means of public record."
- The *Self* mailing informs "recipients that who votes is public information and [lists] the recent voting record of each registered voter in the household. The word 'Voted' appears by names of registered voters who actually voted in the 2004 primary election and the 2004 general election, and blank space appears if they did not vote." The mailing concludes by informing voters that "after the primary election 'we intend to mail an updated chart,' filling in whether the recipient voted in the August 2006 primary."
- The *Neighbors* mailing lists "not only the household's voting records but also the voting records of those living nearby." This mailing also concludes by informing the voter that "we intend to mail an updated chart."

The results of this experiment, shown in Figure 1, have proved to be of great interest to researchers and campaigns alike. While the authors had expected a large increase from the self and neighbors mailings, the magnitude of this effect was much larger than anticipated. Furthermore, replications of this experiment have found similarly large effects for the self and neighbors mailings [Davenport, 2008, Enos, 2008, Grose and Russell, 2008, Mann, 2009, Panagopoulos, 2008]. More importantly, this research has raised critical questions about the role of social pressure in the decision to vote.

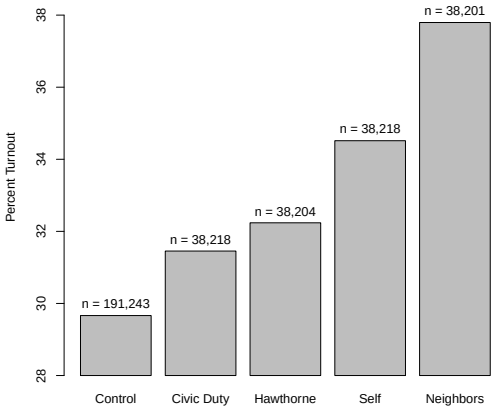

Figure 1: Sample means of voter turnout for control group and four treatment groups in the August 2006 primary election [Gerber et al., 2008, Table 2].

3.2 Theoretical Foundations

The theoretical foundation of the Gerber et al. [2008] experiment is the "calculus of voting" model, due to Downs [1957] and Riker and Ordershook [1968]. The basic theory is that a citizen votes if

$$pB + D > C$$

where p is the probability the vote is pivotal, B is the benefit to the citizen due to the new candidate or her policies, D is the direct benefit of voting, and C is the cost. Gerber et al. argue that the decision to vote can be re-written as

$$pB + \beta_1 D_I + \alpha\pi_r + \beta_3 \pi_r D_I > C$$

where α and β_i are constants, D_I is the intrinsic benefit of voting, and π_r is the perceived probability others learn whether one voted. The term $\beta_3 \pi_r D_I$ is included "to capture the possibility of an interaction between the intrinsic and extrinsic components of civic duty." A positive β_3 implies that a citizen becomes more likely to vote as the perceived probability others learn of the vote increases. A negative β_3 implies that "publicizing one's voting behavior will undercut a citizen's intrinsic motivation, possibly to the point where the citizen becomes even less likely to vote than in the absence of such publicity."

Gerber et al. tested this hypothesis using an "individual's voting propensity as a proxy for the extent to which he or she feels an obligation to vote." They divided the observations into "six subsamples based on the number of votes cast in five prior elections" and further subdivisions by household size and other covariates. They subsequently "conducted a series of logistic regressions and examined the

treatment effects across subgroups," concluding that "the treatment effects on underlying vote propensity are more or less constant, regardless of whether the target group votes often or rarely. [The authors] infer, therefore, that there are no appreciable interactions between social pressure and one's sense of civic duty." In other words, Gerber et al. conclude that the interaction parameter β_3 "appears to be zero," which is especially surprising given the experiment's high statistical power.

3.3 CATE by 2004 Primary Vote

We now re-analyze the Gerber et al. experiment using the framework of the conditional average treatment effect. Prior to investigating treatment effect heterogeneity, we pre-process the data by excluding voters with perfect vote histories, approximately 4% of the sample. We expect such voters to exhibit a strong ceiling effect, and preliminary analysis shows that these voters indeed turnout at a rate over 65%, more than twice the average rate of 32%. In the end, including these voters affects the subsequent results only slightly, but complicates the overall discussion. Since the purpose of this analysis is to argue that treatment effect heterogeneity exists rather than argue for a specific functional form, we exclude this four percent for the sake of clarity.

With this caveat, we begin with a simple hypothesis based on the self and neighbors mailings. Since all individuals in the experiment voted in the November 2004 general election, the only difference in the content of the mailings containing vote history is whether the recipient voted in the August 2004 primary election. The subset that voted in both the August and November 2004 elections had "`Voted  Voted`" next to their names while those who only voted in November 2004 had a blank space plus `Voted` (see Appendix). If the interaction between intrinsic motivation and public knowledge of one's vote, β_3, is positive, then we expect a greater treatment effect among those who voted in the 2004 primary. We can test this by creating a dummy variable, `p2004`, which equals one if the person voted in the 2004 primary, about 40% of the sample, and zero otherwise. To estimate the CATE for 2004 primary vote, $\tau(\texttt{p2004})$, we first re-create the main effects regression in the original paper:

$$Y_i = \alpha + \tau_{\text{civ}} D_{\text{civ}} + \tau_{\text{haw}} D_{\text{haw}} + \tau_{\text{self}} D_{\text{self}} + \tau_{\text{nbr}} D_{\text{nbr}} + \sum_j \beta_j x_{i,j} + \epsilon_i$$

where D_{treat} are dummy variables for each of the four treatments and τ_{treat} are the corresponding estimates of ATE. We separately estimate the CATE for each treatment,

$$\begin{aligned} Y_i = \alpha \quad &+ \quad (\tau_{\text{civ}} + \delta_{\text{civ}}\,\texttt{p2004}) D_{\text{civ}} + (\tau_{\text{haw}} + \delta_{\text{haw}}\,\texttt{p2004}) D_{\text{haw}} + (\tau_{\text{self}} + \delta_{\text{self}}\,\texttt{p2004}) D_{\text{self}} \\ &+ \quad (\tau_{\text{nbr}} + \delta_{\text{nbr}}\,\texttt{p2004}) D_{\text{nbr}} + \gamma\,\texttt{p2004} + \sum_j \beta_j x_{i,j} + \epsilon_i \end{aligned}$$

The results of this regression are shown in specification (b) of Table 1. Although it is necessary to be cautious while interpreting p-values from regressions on large data sets, the treatment effect heterogeneity for `p2004` is strong and significant for both the self and neighbors treatments, but not the civic duty and Hawthorne mailings. In other words, conditioning the treatment effect by `p2004` is a useful distinction for individual treatment effects with the self and neighbors treatments but the data do not support such distinctions for the civic duty and Hawthorne treatments. This interaction is represented graphically in Figure 2, with the CATEs estimated directly using the alternative regression specification in (2).

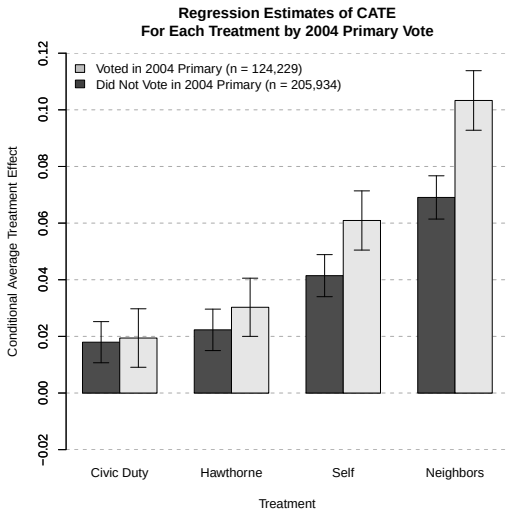

Figure 2: Conditional average treatment effects by 2004 primary vote. Ninety-five percent error bars are determined from the regression standard errors. Results are confirmed by sample estimates (not shown). This regression controls for age, age^2, and household size.

3.4 CATE by Any Previous Vote

A related hypothesis is that those who generally do not vote (except in November 2004) respond differently to treatment compared to those who vote at least on occassion. As above, we create a dummy variable, `anyvote`, which equals one if the individual voted at least once in either the general or primary elections in 2000 or 2002 or the 2004 primary, about 94% of the sample, and zero otherwise. The regression results are shown in specification (c) of Table 1 and graphically in Figure 3. In this case, the treatment effect heterogeneity is dramatic, especially for the neighbors treatment. The subset that generally does not vote receives a 2.4% treatment effect from the neighbors treatment, compared to an 8.5% treatment effect for the subset that votes—more than a three-fold increase.

Table 1: Heterogeneous Treatment Effects by Subgroup in Gerber et al. [2008]

	(a)	(b)	(c)
Civic Duty Treatment	.019** (.003)	.018** (.004)	-.004 (.006)
Hawthorne Treatment	.025** (.003)	.022** (.004)	.001 (.007)
Self Treatment	.049** (.003)	.041** (.004)	.013* (.007)
Neighbors Treatment	.082** (.003)	.069** (.004)	.024** (.007)
Voted in 2004 Primary	-	.135** (.003)	-
Voted in 2004 Primary × Civic Duty	-	.001 (.006)	-
Voted in 2004 Primary × Hawthorne	-	.008 (.006)	-
Voted in 2004 Primary × Self	-	.019* (.006)	-
Voted in 2004 Primary × Neighbors	-	.034** (.006)	-
Any Previous Vote	-	-	.126** (.004)
Any Previous Vote × Civic Duty	-	-	.025** (.007)
Any Previous Vote × Hawthorne	-	-	.027** (.007)
Any Previous Vote × Self	-	-	.038** (.007)
Any Previous Vote × Neighbors	-	-	.062** (.007)
Intercept	-.073** (.007)	-.071** (.007)	-.017** (.008)
N of individuals	330,163	330,163	330,163
Controls	Yes	Yes	Yes
Other Vote History	Yes	Yes	No
R^2	0.06	0.06	0.03

Note: Specification (a) is taken from table 3 of Gerber et al. [2008]. Following their model, "robust cluster standard errors acount for the clustering of individuals within household, which was the unit of assignment." *Controls* are `age`, $\texttt{age}^2$, and an indicator for household size. *Other Vote History* is a list of dummy variables for both primary and general elections in 2000 and 2002 and the 2004 primary. Specification (c) does not include other vote history since the interaction term is a function of the complete vote history.
$^{*}p < 0.05$; $^{**}p < 0.001$

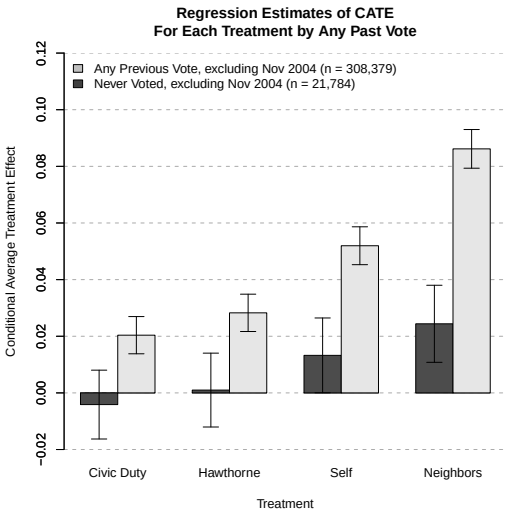

Figure 3: Conditional average treatment effects by any previous vote (excluding November, 2004). Ninety-five percent error bars are determined from the regression standard errors. Results are confirmed by sample estimates (not shown). This regression controls for age, age^2, and household size.

3.5 CATE by Number of Past Votes

Based on this preliminary analysis, we examine the more general interaction between past voting behavior and treatment effect. We further explore this question by following Gerber et al.'s suggestion to create an index based on the number of votes cast in the five prior elections. This quantity is a "proxy for the extent to which [an individual] feels an obligation to vote," represented by the theoretical variable, D_I, which increases as the total number of previous votes increases. Additionally, the "perceived probability others learn whether one voted," π_r, is much greater for the Hawthorne, self, and neighbors mailings over the control group, but is not greater for the civic duty mailing. As such, the observed interaction effect between number of past votes and treatment effect is a suitable empirical test for the theoretical interaction term, $\beta_3 \pi_r D_I$. The coefficient β_3 is considered positive (negative) if the interaction terms for the Hawthorne, self, and neighbors mailings with total past votes are positive (negative). Additionally, since π_r increases across the Hawthorne $<$ self $<$ neighbors treatments, we would expect any effect to be largest (in magnitude) for the neighbors and smallest for the Hawthorne mailing.

We begin with sample estimates of the conditional average treatment effect for each subdivision by vote history. In other words, we select all individuals with `Past Votes` $= 2$, say, and find the difference between the mean level of turnout in a given treatment group minus the mean level of turnout in the control group. These simple differences are shown graphically in Figure 4. While there does not appear to be any trend for the civic duty treatment, there are positive, linear trends for the Hawthorne, self, and neighbors mailings. We next examine this pattern in the regression framework of equation (2) and derive separate CATE estimates for each treatment by past vote history interaction. The results are shown in

Figure 5. Again, there does not appear to be any trend in conditional effect for the civic duty treatment, but there are positive, linear effects for the Hawthorne, self, and neighbors treatments. Additionally, the slope of the linear effect appears to increase from Hawthorne to self to neighbors.

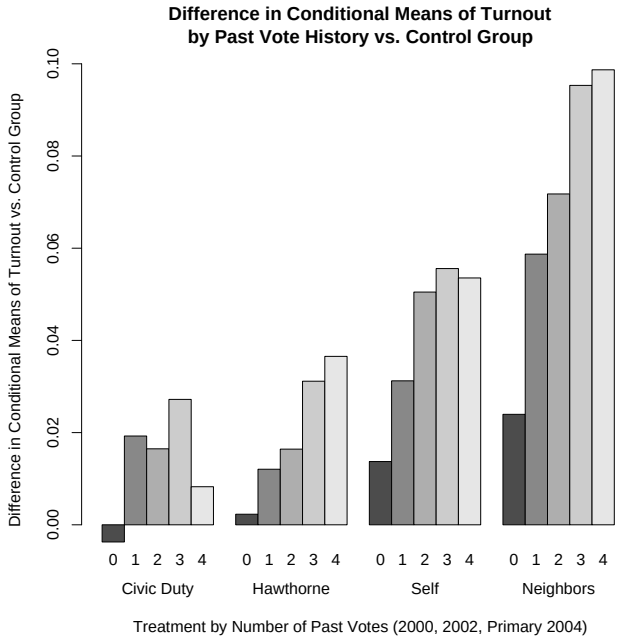

Figure 4: Sample means of conditional average treatment effects for each treatment by past vote history. These values are calculated by subtracting the conditional mean turnout among the control group from the conditional mean turnout for each treatment group.

We can formally test whether there is indeed a linear interaction effect for past vote history with treatment using regression specification (3). The results of this regression are shown in Table 2. As expected, there is no linear interaction effect for the civic duty treatment, but there are strongly significant linear interaction effects for the Hawthorne, self, and neighbors treatments. Interestingly, both the Hawthorne and civic duty mailings have similar interaction effects, although it is unlikely there is enough power to detect a meaningful difference between them. The size of the neighbors interaction effect is especially striking, predicting a linear increase from a 3.6% effect for zero previous votes to a 10.8% effect for four previous votes.

The results in Table 2 provide convincing empirical evidence that the theoretical coefficient β_3 is positive, supporting Gerber et al.'s hypothesis that social pressure "reinforces existing motivation to participate". First, the Hawthorne, self, and neighbors treatments each have positive, significant interactions with the number of past votes. Second, this interaction increases as π_r increases across treatments. We do not extend this interpretation to those voters with perfect vote histories.

It is important to note that in their original analysis, Gerber et al. make the assumption that an

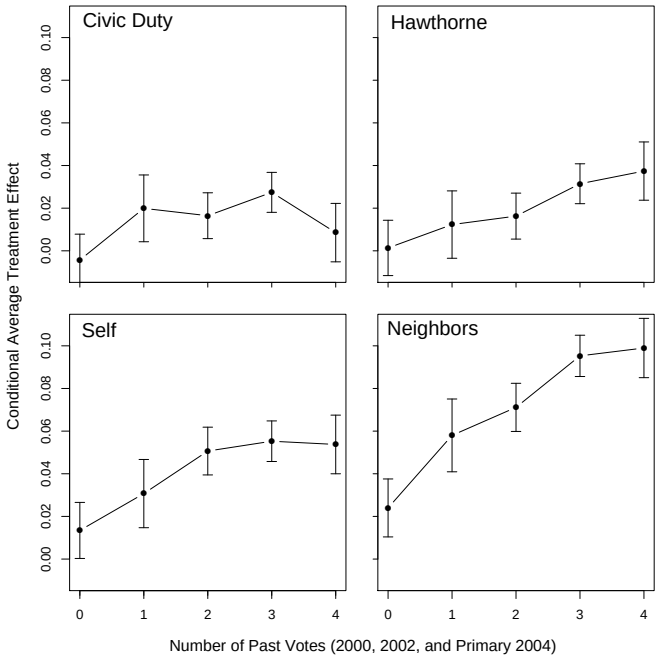

Figure 5: Conditional average treatment effects by vote history for each treatment estimated using regression. Ninety-five percent error bars are determined from the regression standard errors. These regressions control for age, age^2, and household size.

individual's latent vote propensity follows a logistic distribution, such that the "treatments are most effective in terms of votes produced when directed at people with middling vote probabilities." In other words, this assumption imposes the logistic functional form on any treatment effect heterogeneity. As such, if logistic regression does not yield evidence of heterogeneous treatment effects, this null finding does not mean that there is no evidence of heterogeneity but rather that any heterogeneity is indistinguishable from a logistic function. Since the sample estimates in Figure 4 and the corresponding linear regression results show strong evidence of treatment effect heterogeneity, we check our findings using logistic regression. The results (not shown) show signficant quadratic treatment effect heterogeneity by vote history, especially for the self and neighbors mailings, with larger effects at the center of the distribution. The presence of a non-linear effect rather than the linear effect observed above is expected due to the logistic transform. Overall, these results suggest that a logistic distribution only captures some of the heterogeneity, which conforms with recent research on vote propensity and GOTV [Arceneaux and Nickerson, 2009, Hillygus, 2005, Niven, 2004]. For this and other reasons discussed above, we advocate use of the Linear Probability Model, which is agnostic as to the underlying distribution and can be used to measure such heterogeneity directly [Angrist, 2001, Angrist and Pischke, 2009].

In the end, although our results directly contradict the empirical finding in Gerber et al. [2008] that "there are no appreciable interactions between social pressure and one's sense of civic duty," we be-

Table 2: Continuous Heterogeneous Treatment Effects in Gerber et al. [2008]

	(a)	(b)
Civic Duty Treatment	.018** (.003)	.013* (.006)
Hawthorne Treatment	.025** (.003)	.0003 (.006)
Self Treatment	.049** (.003)	.027** (.006)
Neighbors Treatment	.082** (.003)	.036** (.006)
Total Previous Votes	.082** (.001)	.078** (.001)
Total Previous Votes × Civic Duty	-	.002 (.002)
Total Previous Votes × Hawthorne	-	.010** (.002)
Total Previous Votes × Self	-	.009** (.002)
Total Previous Votes × Neighbors	-	.018** (.002)
Intercept	.008 (.007)	.019* (.008)
N of individuals	330,163	330,163
Controls	Yes	Yes
Other Vote History	No	No
R^2	0.05	0.05

Note: Specification (a) is taken from table 3 of Gerber et al. [2008]. Following their model, "robust cluster standard errors acount for the clustering of individuals within household, which was the unit of assignment." Voters with with five previous votes (4% of the sample) were excluded from the analysis due to significantly higher turnout and reduced statistical power. *Controls* are `age`, $\texttt{age}^2$, and an indicator for household size.
$^{*}p < 0.05$; $^{**}p < 0.001$

lieve that the observed treatment effect heterogeneity significantly strengthens Gerber et al.'s theoretical argument. All together, this re-analysis demonstrates the practical and theoretical importance of exploring heterogeneous treatment effects, as a standard "toplines" approach could not detect the significant differences in treatment effects across voter subgroups.

4 Rock the Vote Text Message Experiment

4.1 Experimental Setup

To illustrate non-linear treatment effect heterogeneity, we analyze a randomized experiment conducted in the 2008 election by Rock the Vote, a non-partisan youth advocacy and mobilization group [Kennedy and Mayorga, 2008]. In this election cycle, Rock the Vote executed a major voter registration drive, eventually registering more than two million 18 to 29 year-olds. One key feature of this initiative was an online voter registration tool used to register over one million voters. Of those that used the online system, approximately 200,000 opted-in to receive text message updates from Rock the Vote. The low cost and ease of mass text messaging makes this group ideal for experimentation, and Rock the Vote conducted over twenty randomized experiments over the course of the election cycle on this and

other groups. We focus on an experiment from the Indiana and Pennsylvania 2008 Presidential primary elections, originally described by Kennedy and Mayorga [2008].

As part of their voter registration efforts, Rock the Vote sought to maximize the proportion of downloaded voter registration forms that individuals successfully completed. One strategy was to send text messages encouraging individuals to mail in their registrations and reminding them of their state's registration deadline. An example text for Pennsylvania reads: "`Are you registered to vote? If not, do it at rockthevote.com/PA or call 877-868-3772 to check status. PA's deadline is this Monday. Please fwd.`" Since voter databases are only updated periodically, reminders were sent to all individuals who had opted-in for text messages and had not appeared on the most recent voter registration list, meaning that many of the individuals who recieved the reminder had already registered. A sample from this population was randomly assigned to the treatment or control group,[10] with 4,156 individuals assigned to treatment and 2,104 individuals assigned to control. Randomization checks (not shown) confirm that treatment assignment does not depend on the observed covariates. The observed outcome is whether or not each individual has registered to vote by the deadline.

One important factor is the lag between the date of download and the text message reminder, which is measured in days.[11] We hypothesize that those individuals who downloaded the registration form several months earlier are much more likely to have already sent in their registration forms than those who downloaded the form on the previous day. At the same time, we believe that a registration reminder is more likely to be effective among recent downloads than among those for whom downloading the registration form is not as immediate. In both cases, while we can posit a general trend, the functional forms of the baseline and interaction effects are unknown.

Several studies have analyzed the effects of timing on the effectiveness of voter mobilization and suggest that interventions closer to election day are generally more effective [Nickerson, 2006, Green and Gerber, 2008]. While there are no prior studies on text messaging and registration, Dale and Strauss [2007] conduct an experiment in which text messages increase turnout by 3.1 percent (ITT estimate) [see also Suárez, 2006, Panagopoulos, 2009].

4.2 Results

Basic sample estimates show that text message reminders have a small but significant effect on registration: 64.7% of the treatment group registered to vote, compared to 60.3% of the control group. This corresponds to a 4.4 percentage point average treatment effect, close to the result from Dale and Strauss

[10] The treatment group was further divided to receive one of four closely related messages. The results across the four messages are statistically indistinguishable and are treated as a single group in this analysis.

[11] Due to sparsity and extreme values, individuals with a lag greater than 200 days, approximately 2% of the data set, were excluded. The results are nearly identical to those produced here, but the graphical representation is less clear.

[2007], which is confirmed by the simple regression estimate shown in specification (a) of Table 3. We subsequently check this estimate while controlling for several covariates: age, gender, home state, and days since download. Since we have no prior beliefs about functional forms for age or days since download, we fit a semiparametric generalized additive model, allowing both continuous covariates to have non-linear effects on registration.[12] The results of this regression are shown in Table 3 and the baseline effects of age and days since download on registration are shown in Figure 6. As demonstrated in the figure, the baseline effect of age on registration is clearly non-linear, with comparatively low registration rates among 18 to 20 year-olds, a sharp increase in registration between 20 and 23, and a relatively level positive rate from 23 to 29. The baseline effect of days since registration is low for recent downloads, with the registration rate increasing until approximately 90 days since download. This result confirms the hypothesis that individuals who downloaded the registration form much earlier were more likely to have already completed the registration process than otherwise.

The regression results in Table 3 report the effective degrees of freedom for the smooth terms and the p-value of the corresponding F-test. In this case, the p-values for m(age) and m(days since download) are highly significant at $p = 2 \times 10^{-12}$ and $p = 1 \times 10^{-15}$ respectively. The strongly non-linear term for age uses nearly 4 effective degrees of freedom, while the approximtely quadratic term for days since download uses roughly 2 effective degrees of freedom. We can use ANOVA to test whether these smooth terms should be replaced by parametric terms: including a linear or quadratic term for age is significantly worse than using the smooth age term ($p = 0.01$ and $p = 0.02$ respectively); including a linear term for days since download is also significantly worse than using the smooth term ($p = 8 \times 10^{-5}$).[13]

We can extend the semiparametric analysis to determine whether there is significant treatment effect heterogeneity, testing both age and days since download for non-linear interactions with treatment. The interaction effect for age is not significant, with $p = 0.71$, but the interaction effect for days since registration, shown in Table 3, is strongly significant at $p = 0.006$.[14] The conditional average treatment effect for days since download is shown in Figure 7. Like the corresponding figures for the social pressure experiment, the CATE plot shows the average treatment effect at each value of days since download, $\hat{m}$(days since download) $+ \hat{\tau}$, where the standard errors have been estimated as in Equation (2). The CATE plot clearly demonstrates that the treatment has a constant positive effect on registration between approximately 0 and 90 days since download, decreasing quickly as the lag since download increases.

[12] The smooth terms are estimated using *thin plate regression splines* in the package `mgcv`. Different levels of smoothing were tried, but differences in the final results proved minor overall. Other smoothing methods, especially cubic regression splines, also produce similar results.

[13] Since the smooth term for days since download uses very close to 2 effective degrees of freedom, the F-test between the smooth term and a quadratic approximation is indeterminate. A simple comparison shows that the model using the non-linear term has a lower residual variance than the corresponding parametric model.

[14] We also confirmed that there are no significant interactions between the treatment and the dichotomous covariates for gender and home state ($p = 0.76$ and $p = 0.59$).

Table 3: Semiparametric Regression for Rock the Vote's Text Message Reminders

	(a)	(b)	(c)
Parametric Coefficients:			
Treatment	.044** (.012)	.043** (.013)	.052** (.016)
Intercept	.603** (.011)	.394** (.018)	.394** (.017)
Nonparametric Coefficients:		(effective degrees of freedom)	
m(Age)	-	3.82**	3.83**
m(Days Since Download)	-	1.97**	1.50**
m(Days Since Download × Treatment)	-	-	2.19**
N of individuals	6,260	6,260	6,260
Controls	No	Yes	Yes
(Adj.) R^2	0.002	0.05	0.05

Note: Specification (a) is the simple OLS regression estimate for the causal effect and does not contain any nonparametric terms. Specification (b) adds controls of home state and gender in addition to non-linear terms for age and days since download. Specification (c) includes a non-linear interaction effect for days since download, as specified in equation (4).
$^{*}p < 0.05$; $^{**}p < 0.01$

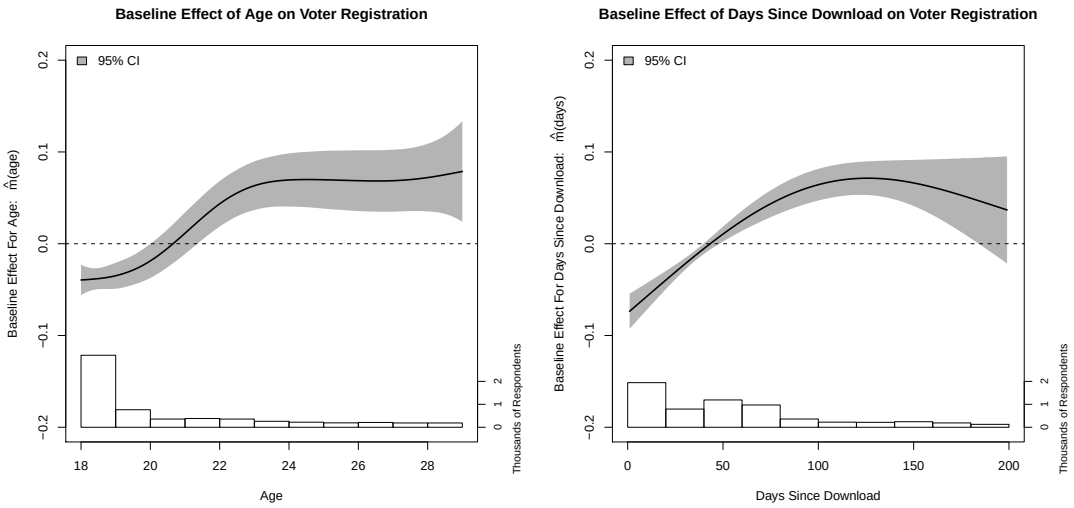

Figure 6: The baseline effect of age and days since download on voter registration, estimated using splines with a generalized additive model. The distribution of each variable is shown in a histogram superimposed at the bottom of the plot. The semiparametric regression also controls for home state and gender.

ANOVA confirms that the smooth interaction term should not be replaced by a linear term, with $p = 0.03$. Since the estimate shown in the CATE plot controls for the baseline effects in Figure 6, the non-linear interaction effect suggests that registration reminders are indeed effective at getting individuals to register to vote, but that this effect is only significant for those individuals who downloaded the registration form relatively recently.

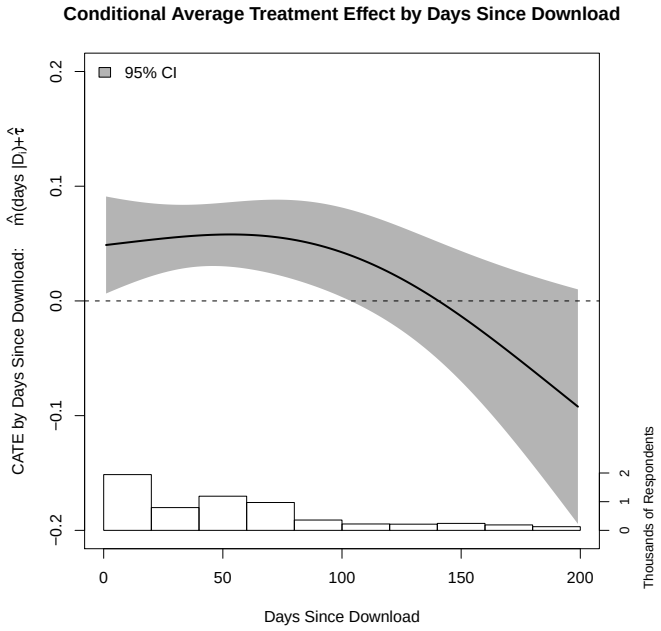

Figure 7: The conditional average treatment effect on voter registration by days since download, estimated using splines with a generalized additive model. The distribution of days since download is shown in a histogram superimposed at the bottom of the plot.

In the end, neither basic topline approaches nor standard regression techniques could identify the treatment effect heterogeneity described here. The Rock the Vote experiment, however, is just one of many examples with continuous covariates. In such cases, unless there is a strong belief in a functional form *a priori*, GAMs are very useful modeling tools.

5 Discussion

Correctly estimating the average treatment effect is of critical importance in randomized experiments. In many situations where this is readily calculated, however, only estimating ATE can hide important empirical and theoretical results. In such cases, the conditional average treatment effect is a simple but powerful approach for identifying key differences in treatment effects across covariates. The use of linear regression and generalized additive models, especially, gives analysts a practical and flexible framework for estimating conditional treatment effects in a broad range of situations. The re-analysis

of Gerber et al. [2008] demonstrates that even a straightforward linear regression framework can reveal important and previously undetected variation in the experimental results. The analysis of the Rock the Vote experiment similarly shows the benefits of non-linear modeling for heterogeneous treatment effects. For both examples, a simple "toplines" approach would have missed important variation in experimental results. Furthermore, an *ad hoc* subgroup analysis would have run the risk of overfitting the data, leading to potentially erroneous conclusions.

There are many potential extensions to the methods presented here. One extension is to the analysis of experiments matched by propensity score. Although more research on the behavior of such estimators is required, both linear regression and generalized additive models are readily adapted to allow for observation weights via the inverse probability weighting estimator [Hirano et al., 2003]. Another important extension is for noncompliance and nonresponse, especially incorporating local average treatment effects [Imbens and Angrist, 1994] into the semiparametric framework of GAMs.

Throughout this paper we have presented a Frequentist approach for the analysis of randomized experiments. It seems to us that substantial advantages can be gained by careful modeling within a Bayesian framework, as outlined by Imbens and Rubin [1997], which also provides a powerful method for addressing noncompliance. This can be further expanded to include Bayesian methods for estimating non-linear effects [Ruppert et al., 2003]. A Bayesian approach can also minimize problems of "data snooping" associated that are common with a large number of potential covariates and no clear hypotheses about interactions *ex ante*. Dixon and Simon [1991], for example, use Bayesian regression with a regularization prior over all possible interactions to identify significant subgroup interactions. This approach, however, is limited to categorical covariates and is not easily expanded to the continuous case. Another possibility is to abandon parametric and semiparametric frameworks entirely and use the nonparametric Bayesian Additive Regression Trees (BART), which has recently been extended for causal inference applications [Hill and McCulloch, 2008]. While this algorithmic approach shows promise as an exploratory technique, the more "hands on" GAM framework is preferable in most descriptive rather than predictive situations.

In the end, heterogeneous treatment effects matter. A wide range of statistical techniques can be used to estimate this heterogeneity, either using methods proposed here or elsewhere. As such, the careful analysis of heterogeneous treatment effects should be a standard component of any experimental results, whether in political research or in campaign work. This methodological development is necessary for the growing number of political science experiments to continue to advance our understanding of political participation.

References

J. Aitchison. Statistical problems of treatment allocation. *Journal of the Royal Statistical Society, Series A*, 133:206–238, 1970.

J. D. Angrist. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business and Economic Statistics*, 19(2):2–28, 2001.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

J. D. Angrist, G. W. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

K. Arceneaux and D. W. Nickerson. Who is mobilized to vote? a re-analysis of eleven randomized field experiments. Unpublished manuscript, 2009.

S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355:1064–1069, 2000.

N. Beck and S. Jackman. Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 42(2):596–627, April 1998.

A. Björklund and R. Moffitt. The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics*, 69(1):42–49, February 1987.

D. P. Byar. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4:255–263, 1985.

D. P. Byar and D. K. Corle. Selecting optimal treatment in clinical trials using covariate information. *Journal of Chronic Diseases*, 30:445–459, 1977.

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

D. R. Cox. Interaction. *International Statistical Review*, 52(1):1–24, April 1984.

L. J. Cronbach and R. E. Snow. *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington, New York, 1977.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, August 2008.

R. V. Crump, V. J. Hotz, G. Imbens, and O. Mutnik. Dealing with limited overlap in estimatino of average treatment effects. *Biometrika*, forthcoming.

A. Dale and A. Strauss. Text messaging as a youth mobilization tool: An experiment with a post-treatment survey. Presentation at the Annual Meeting of the Midwest Political Science Association, April 2007.

T. Davenport. Public accountability and participation: The effects of a feedback intervention on voter turnout in a low salience election. Unpublished manuscript, Yale University, 2008.

R. H. Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125:141–173, 2005.

D. O. Dixon and R. Simon. Bayesian subset anaylsis. *Biometrics*, 47(3):871–881, September 1991.

A. Downs. *An Economic Theory of Democracy*. Harper and Row, New York, 1957.

J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. The growth and development of experimental research in political science. *American Political Science Review*, 100(4):627–635, November 2006.

R. D. Enos. The effect of geography-based group threat on voter mobilization: A field experiment. Working Paper, University of California, Los Angeles, 2008.

R. L. Eubank, J. D. Hart, D. G. Simpson, and L. A. Stefanski. Testing for additivity in non-parametric regression. *Annals of Statistics*, 23:1896–1920, 1995.

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

C. E. Frangakis and D. B. Rubin. Principal stratification and causal inference. *Biometrics*, 58:21–29, March 2002.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, 23(2):237–249, 2008a.

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193, 2008b.

A. Gelman. A statisticians perspective on "mostly harmless econometrics: An empiricists companion", by joshua d. angrist and jörn-steffen pischke. *The Stata Journal*, 9(1):1–6, 2009.

A. Gelman and Z. Huang. Estimating incumbancy advantage and its variation, as an example of a before-after study. *Journal of the American Statistical Association*, 103(482):437–446, 2008.

A. Gelman and G. King. Enhancing democracy through legislative redistricting. *American Political Science Review*, 88(3):541–559, 1994.

A. S. Gerber and D. P. Green. Field experiments and natural experiments. In J. M. Box-Steffensmeier, H. E. Brady, and D. Collier, editors, *Handbook of Political Methodology*, chapter 15, pages 357–381. Oxford University Press, 2008.

A. S. Gerber, D. P. Green, and C. W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48, February 2008.

D. P. Green and A. S. Gerber. *Get Out The Vote: How to Increase Voter Turnout (Second Edition)*. Brookings Institution Press, 2008.

C. R. Grose and C. A. Russell. Avoiding the vote: A theory and field experiment of the social costs of public political participation. Available at SSRN: http://ssrn.com/abstract=1310868, 2008.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

B. B. Hansen and J. Bowers. Attributing effects to a cluster randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, Forthcoming, 2009.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

J. J. Heckman. Microdata, heterogeneity and the evaluation of public policy: Nobel lecture. *Journal of Political Economy*, 109(4):673–748, August 2001.

J. J. Heckman. Econometric causality. *IZA Discussion Paper Series*, (3425), March 2008.

J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, May 2005.

J. L. Hill and R. E. McCulloch. Bayesian nonparametric modeling for causal inference. NYU Technical Paper, November 2008.

D. S. Hillygus. Campaign effects and the dynamics of turnout intention in election 2000. *Journal of Politics*, 67(1):50–68, 2005.

K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, July 2003.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960, December 1986.

Y. Horiuchi, K. Imai, and N. Taniguchi. Designing and analyzing randomized experiments: Application to a japanese election survey experiment. *American Journal of Political Science*, 51(3):669–687, July 2007.

S. Hunsberger and D. A. Follmann. Testing for treatment and interaction effects in semi-parametric analysis of covariance. *Statistics in Medicine*, 20:1–19, 2001.

K. Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, May 2005.

K. Imai and A. Strauss. Planning the optimal get-out-the-vote campaign using randomized field experiments. *Working Paper*, 2009.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, March 1994.

G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25:305–327, 1997.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009.

P. O. Johnson and L. C. Fay. The johnson-neyman technique, its theory and application. *Psychometrika*, 15(4):349–367, December 1950.

P. O. Johnson and J. Neyman. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, 1:57–93, 1936.

C. Kennedy and M. Mayorga. Text message experiments in 2008. Available at `http://www.rockthevote.com/assets/publications/research/text-message-experiments-2008.pdf`, November 2008.

R. J. A. Little. Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012, 1993.

C. B. Mann. Assessing the risks of social pressure for voter mobilization: A large scale field experiment. Unpublished manuscript, Yale University, 2009.

P. McCullagh and J. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 1989.

J. Neyman. Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society II*, 2:107–180, 1923.

D. W. Nickerson. Forget me not? the importance of timing in voter mobilization. Presentation at the annual meeting of the American Political Science Association, September 2006.

D. Niven. The mobilization solution? face-to-face contact and voter turnout in municipal elections. *Journal of Politics*, 66:868–885, 2004.

C. Panagopoulos. Assessing the risks of social pressure for voter mobilization: A large scale field experiment. Institution for Social and Policy Studies 40th Anniversary Conference, Yale University, November 14-15, 2008, New Haven, CT, 2008.

C. Panagopoulos, editor. *Politicking Online: The Transformation of Election Campaign Commuications*. New Brunswick, NJ: Rutgers University Press, 2009.

J. Pearl. *Causality*. Cambridge University Press, 2000.

S. J. Pocock, S. E. Assmann, L. E. Enos, and L. E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:2917–2930, 2002.

R. F. Potthoff. On the johnson-neyman technique and some extensions thereof. *Psychometrika*, 29(3): 241–256, 1964.

W. H. Riker and P. C. Ordershook. A thoery of the calculus of voting. *American Political Science Review*, 62:25–43, March 1968.

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

P. M. Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, 365:176–186, 2005.

D. B. Rubin. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, 47(1):1–52, 1985.

S. L. Suárez. Mobile democracy: Text messages, voter turnout and the 2004 spanish general election. *Representation*, 42(2):117–128, 2006.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall, 2006.

Figure 8: Civic Duty Mailing [Gerber et al., 2008]

30426-2 ||| || || || | XXX

For more information: (517) 351-1975
email: etov@grebner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRSRT STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C002
THE JONES FAMILY
9999 WILLIAMS RD
FLINT MI 48507

Dear Registered Voter:

DO YOUR CIVIC DUTY AND VOTE!

Why do so many people fail to vote? We've been talking about this problem for years, but it only seems to get worse.

The whole point of democracy is that citizens are active participants in government; that we have a voice in government. Your voice starts with your vote. On August 8, remember your rights and responsibilities as a citizen. Remember to vote.

DO YOUR CIVIC DUTY - VOTE!

Figure 9: Hawthorne Mailing [Gerber et al., 2008]

30424-1 ||| || || ||

For more information: (517) 351-1975
email: etov@grebner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRSRT STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C001
THE SMITH FAMILY
9999 PARK LANE
FLINT MI 48507

Dear Registered Voter:

YOU ARE BEING STUDIED!

Why do so many people fail to vote? We've been talking about this problem for years, but it only seems to get worse.

This year, we're trying to figure out why people do or do not vote. We'll be studying voter turnout in the August 8 primary election.

Our analysis will be based on public records, so you will not be contacted again or disturbed in any way. Anything we learn about your voting or not voting will remain confidential and will not be disclosed to anyone else.

DO YOUR CIVIC DUTY - VOTE!

Figure 10: Self Mailing [Gerber et al., 2008]

30422-4 ||| || || | ||

For more information: (517) 351-1975
email: etov@grebner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRSRT STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C050
THE WAYNE FAMILY
9999 OAK ST
FLINT MI 48507

Dear Registered Voter:

WHO VOTES IS PUBLIC INFORMATION!

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse.

This year, we're taking a different approach. We are reminding people that who votes is a matter of public record.

The chart shows your name from the list of registered voters, showing past votes, as well as an empty box which we will fill in to show whether you vote in the August 8 primary election. We intend to mail you an updated chart when we have that information.

We will leave the box blank if you do not vote.

DO YOUR CIVIC DUTY - VOTE!

OAK ST	Aug 04	Nov 04	Aug 06
9999 ROBERT SMITH		Voted	_____
9999 LAURA BETH	Voted	Voted	_____

Figure 11: Neighbors Mailing [Gerber et al., 2008]

30423-3 ||| || || | |||

For more information: (517) 351-1975
email: etov@grebner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRSRT STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C050
THE JACKSON FAMILY
9999 MAPLE DR
FLINT MI 48507

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY - VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____
9999 BRIAN JOSEPH JACKSON		Voted	_____
9991 JENNIFER KAY THOMPSON		Voted	_____
9991 BOB R THOMPSON		Voted	_____
9993 BILL S SMITH			_____
9989 WILLIAM LUKE CASPER		Voted	_____
9989 JENNIFER SUE CASPER		Voted	_____
9987 MARIA S JOHNSON	Voted	Voted	_____
9987 TOM JACK JOHNSON	Voted	Voted	_____
9987 RICHARD TOM JOHNSON		Voted	_____
9985 ROSEMARY S SUE		Voted	_____
9985 KATHRYN L SUE		Voted	_____
9985 HOWARD BEN SUE		Voted	_____
9983 NATHAN CHAD BERG		Voted	_____
9983 CARRIE ANN BERG		Voted	_____
9981 EARL JOEL SMITH			_____
9979 DEBORAH KAY WAYNE		Voted	_____
9979 JOEL R WAYNE		Voted	_____