Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods

Justin Grimmer * Solomon Messing † Sean J. Westwood ‡

July 6, 2013

Abstract

Randomized experiments are increasingly used to study political phenomena because they can credibly estimate the average effect of a treatment on a population of interest. But political scientists are often interested in how effects vary across sub-populations—*heterogeneous* treatment effects —and how differences in the content of the treatment affects responses—the response to *heterogeneous* treatments. Several new methods have been introduced to estimate heterogeneous effects, but it is difficult to know if a method will perform well for a particular data set. Rather than use only one method, we show how an ensemble of methods—weighted averages of estimates from individual models—accurately measure heterogeneous effects. Building on a large literature on ensemble methods, we show the close relationship between out of sample prediction and accurate estimation of heterogeneous treatment effects and demonstrate how pooling models leads to superior performance to individual methods across diverse problems. We apply the ensemble method to two experiments, illuminating how constituents reward and punish legislators for particularistic spending.

*Assistant Professor, Department of Political Science, Stanford University; Encina Hall West 616 Serra St., Stanford, CA, 94305

†Ph.D. candidate, Department of Communication, Stanford University, 450 Serra Mall, Building 120, Room 110, Stanford, CA, 94305

‡Ph.D. candidate, Department of Communication, Stanford University, 450 Serra Mall, Building 120, Room 110, Stanford, CA, 94305

1 Introduction

Experiments are increasingly used to test theories of politics and political conflict (Gerber and Green, 2012). Experiments are used because they provide credible estimates of the effect of an intervention for a sample population. But underlying this average effect for a sample may be substantial variation in how particular respondents respond to treatments: there may be heterogeneous treatment effects. And this variation may provide theoretical insights, revealing how the effect of interventions depend on participants' characteristics. The variation may also be practically useful, providing guidance on how to optimally administer treatments (Imai and Strauss, 2011; Imai and Ratkovic, 2013). And the variation may also be useful for extrapolating the findings of an experiment to a broader population of interest (Hartman et al., 2012). Further scholars are increasingly making use of experimental designs with many conditions, in order to examine how differences in treatment content affects response—the effect of heterogeneous treatments (Hainmueller and Hopkins, 2013; Hainmueller, Hopkins and Yamamoto, 2013).

A growing literature has contributed new methods for estimating *heterogeneous* effects (Hastie, Tibshirani and Friedman, 2001; Imai and Strauss, 2011; Green and Kern, 2012; Hainmueller and Hazlett, 2012; Imai and Ratkovic, 2013). Each of the methods provide new and important insights into how to reliably capture heterogeneity in treatment response. To identify systematic variation in treatment response and to separate it from variation due to simple randomness each of the new methods combines information in the data with necessary and consequential assumptions about the data generating process (Hastie, Tibshirani and Friedman, 2001). While the assumptions are often minimal and designed to maximize a method's flexibility, it is difficult to know before hand if a method's particular assumptions fit any one application well.

Rather than rely on a single method to estimate heterogeneous treatment effects, we show how a weighted average of methods for estimating heterogeneous effects—an ensemble—provides accurate estimates across diverse problems. We build on the ensemble method

super learning (van der Laan, Polley and Hubbard, 2007), using out of sample prediction to weight the contribution of methods to the final estimate of heterogeneous effects and show the close relationship of super learning to other ensemble methods (van der Laan, Polley and Hubbard, 2007; Hillard, Purpura and Wilkerson, 2008; Montgomery, Hollenbach and Ward, 2012). Weighting based on out of sample performance is useful, we show, because methods that predict well out of sample will also accurately estimate heterogeneous effects. Using Monte Carlo simulations we show that the ensemble outperforms constituent methods across diverse problems because the ensemble attaches greater weight to methods that have better estimates of the heterogeneous effects for the particular task at hand.

We apply the ensemble method to examine how constituents evaluate how legislators' claim credit for particularistic spending in the district and criticism of those credit claiming efforts (Mayhew, 1974; Grimmer, Messing and Westwood, 2012). We show that constituents evaluate credit claiming messages based on easily acquired information—such as the type of project the legislator claims credit for, rather than the amount of money allocated to the project. Further, we show that conservative constituents punish legislators for deficit spending, while strong liberals *reward* legislators for deficit spending.

Throughout the paper we explain that ensembles are useful because they are flexible and can be tuned to the particular problem at hand. They are also useful because they ensure that we make full use of impressive methodological innovations in the estimation of heterogeneous treatment effects. As we explain in the conclusion, ensemble methods are best conceived of as a companion to new constituent methods: better individual methods for estimating heterogeneous effects will lead to better ensemble estimates and the ensembles provide a new method for evaluating individual methods for estimating heterogeneous effects.

2 Experiments and Conditional Average Treatment Effects

We follow a large prior literature and formalize the estimation of heterogeneous effects using potential outcome notation (Holland, 1986; Green and Kern, 2012; Imai and Ratkovic, 2013).

Suppose that we have a sample of $\mathcal{N}$, $(i = 1, \ldots, N)$ individuals from a population $\mathcal{P}$ and that participants are randomly assigned to one of $K+1$ conditions. Participant i's condition will be given by T_i, $(T_i \in \{0, 1, 2, \ldots, k\})$. If, for example, $T_i = k$ then individual i was assigned to the k^{th} treatment condition. Denote respondent i's response to condition k with the potential outcome $Y_i(k)$.[1] We will analyze dichotomous dependent variables, though the ensemble methods generalize easily to continuous or other dependent variables.

To measure the effect of an intervention for the entire population of interest, scholars commonly report an Average Treatment Effect (ATE) across two conditions. Consider two treatment arms k and $k^{'}$, and let $\phi(k, k^{'})$ denote the ATE across the two conditions,

$$\phi(k, k^{'}) \quad = \quad E[Y(k) - Y(k^{'})].$$

Randomly assigning participants to arms of the treatment ensures that the $\phi(k, k^{'})$ is identified for each combination of k and $k^{'}$, which is commonly estimated with a difference in means across conditions.

The ATE measures the effect of the intervention over the entire population, but to measure how treatment effects vary across respondent characteristics we will estimate a *conditional* average treatment effect (CATE) (Imai and Strauss, 2011; Green and Kern, 2012; Imai and Ratkovic, 2013). The CATE, measures the average treatment effect for respondents who share a set of characteristics. To formalize this definition, suppose that for each respondent i we collect J covariates $(j = 1, 2, \ldots, J)$, $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$, with values of the covariates collected in the set $\mathcal{X}$. We can then define the CATE for covariate profile $\boldsymbol{x} \in \mathcal{X}$, and treatment arms k and $k^{'}$ as $\phi(k, k^{'}, \boldsymbol{x})$,

$$\phi(k, k^{'}, \boldsymbol{x}) \quad = \quad \mathrm{E}[Y(k) - Y(k^{'}) | \boldsymbol{X} = \boldsymbol{x}]. \tag{2.1}$$

[1]The fundamental problem of causal inference ensures that we observe only one response for each of our respondents. We will also make the usual SUTVA assumptions, which are particularly likely to hold in our survey experiments.

A treatment effect is heterogeneous if the value of Equation 2.1 varies as we consider different strata of participants. As before, random assignment to treatment conditions is sufficient to identify the CATE.

The CATE measures the effect for respondents who share identical values of all J covariates. We may be interested, however, in how responses to the treatment effect vary across a subset of covariates, or a single covariate. Suppose that we are interested in estimating the marginal effect of a subset of covariates $\boldsymbol{X}_S = (X_{s_1}, X_{s_2}, \ldots, X_S)$. Define the *marginal* conditional average treatment effect MCATE for covariates $\boldsymbol{X}_S$ and treatment conditions k and $k^{'}$, $\phi(k, k^{'}, \boldsymbol{x}_S)$,

$$\begin{aligned}
\phi(k, k^{'}, \boldsymbol{x}_S) &= \int \phi\left(k, k^{'}, (X_1, X_2, \ldots, \boldsymbol{X}_S = \boldsymbol{x}_S, \ldots, X_J)\right) dF_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S} \\
&= \int E\left[Y(k) - Y(k^{'})|(X_1, X_2, \ldots, \boldsymbol{X}_S = \boldsymbol{x}_S, \ldots, X_J)\right] dF_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S}
\end{aligned}$$

where $F_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S}$ is the joint distribution of the covariates $\boldsymbol{X}_{-S}$, given that $\boldsymbol{X}_S = \boldsymbol{x}_S$. In words, Equation 2.2 shows that MCATEs are averages of CATEs, where the value of the covariate of interest is fixed and $F_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S}$ is used to weight other covariate profiles. This notation corresponds to well known quantities of interest scholars commonly compute. For example, when applying a parametric bootstrap, such as `Clarify`, it is common to set variables other than the covariate of interest to the sample means (King, Tomz and Wittenberg, 2000). This is equivalent to selecting $F_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S}$ to place all probability on the sample means and zero elsewhere. It may be possible to estimate the joint distribution, facilitating extrapolation from the sample population to some other sample or population (Hartman et al., 2012). And finally, we may average over all other possible covariate values—for discrete random variables—or a grid of values for continuous random variables. This is equivalent to setting $F_{\boldsymbol{X}_{-S}|\boldsymbol{X}_S=\boldsymbol{x}_S}$ to a uniform distribution.

3 Estimation of Heterogeneous Treatment Effects with a Weighted Ensemble of Methods

When there are a large number of participants in each condition and participants who share the same set of covariates, then reliable estimation of ATEs, CATEs, and MCATEs is straightforward. The random assignment of participants to treatments ensures that a difference in means across treatment arms will reliably estimate the ATE and a difference in means across arms among respondents with the same set of covariates provides an accurate estimate of CATEs and MCATEs. With a large number of participants, the differences computed with naive differences in means will tend to reflect systematic differences (Gelman, Hill and Yajima, 2012). But for more heterogeneous treatments with a large number of conditions, or covariates that have few observations who share the exact same covariates, a simple difference in means will be a less reliable estimate of the effect of treatments. When the sample size is relatively small, naïve differences will be likely to reflect random variation in the sample, rather than systematic differences in the underlying methods because there will be few observations who share the exact same characteristics. This renders unusable the usual method for estimating heterogeneous treatment effects: interacting treatment indicators with covariates in a regression.

The goal in estimating heterogeneous effects is to separate the systematic responses from differences solely due to chance of the random assignment. Several new methods provide novel ways to identify the systematic effects. Each method m estimates the *response* surface for any treatment k and covariates $\boldsymbol{x}$,

$$g_m(k, \boldsymbol{x}_S) \quad = \quad E[Y|T = k, \boldsymbol{x}] \tag{3.1}$$

and quantities of interest are computed by taking differences across the response surfaces.

To estimate Equation 3.1 each of the methods vary necessary and consequential assumptions about how treatment assignment and covariates alter the response surface. For example, one approach to estimating heterogeneous treatment effects is to use regression

trees (Imai and Strauss, 2011) and Bayesian Additive Regression Trees (BART) to estimate CATEs and MCATEs (Chipman, George and McCulloch, 2010; Green and Kern, 2012). The trees subdivide the data repeatedly, developing decision rules to split the data to make more accurate predictions. An ensemble of the trees is then used to model the response surface and estimate the heterogeneous effects. Other methods start from a more familiar regression framework and then use data and assumptions to identify systematic differences. For example, LASSO methods use a penalty to shrink some coefficients to zero (Hastie, Tibshirani and Friedman, 2001). Imai and Ratkovic (2013) extend and generalize LASSO, introducing a model that has two different penalties—one for covariates and another for variation in treatment effects. A different, though related, approach is to impose a model that shrinks the coefficients to a common mean, allowing only the strongest coefficients to take on distinct values (Gelman et al., 2008). Hainmueller and Hazlett (2012) extend this idea further and include a much more flexible modeling approach with *Kernel Regularized Least Squares* (KRLS) and prove useful statistical properties of the algorithm.[2] And still other methods balance between the two types of smoothing. Elastic-Net is a method that includes penalties that both shrink to zero and shrink to a common mean, with the weight attached to each penalty determined by a parameter α, $0 < \alpha < 1$ (Hastie, Tibshirani and Friedman, 2001).

Each of the methods have been shown to perform well on important political science examples. Yet, knowing the ideal method to apply in any one experiment requires knowledge about the data generating process that assumes the heterogeneous effect sizes are known—exactly what we set out to estimate. For example, Imai and Ratkovic (2013) impose a sparseness assumption, using a method that identifies a set of treatments and covariates that have no effect on the response surface. In contrast, Hainmueller and Hazlett (2012) assume the estimates are more *dense*, smoothing many of the coefficient estimates to approximately the same value (Hastie, Tibshirani and Friedman, 2001).

The appropriateness of those modeling assumptions will vary across substantive prob-

[2]Both KRLS, Find It, and other methods discussed here have applications that extend well beyond measuring heterogeneous treatment effects

lems. New and diverse methods, then, are essential for estimating accurate heterogeneity in treatment effects. But relying on a single method will result in sub-optimal performance across diverse problems. When the assumptions fit the data generation process, the model will perform well, but when the assumptions are a poor fit the method will perform poorly.

In place of using a single method to estimate heterogeneous effects, we suggest generating an ensemble of estimators. As we show below, we use a weighted ensemble because it attaches the greatest weight to the methods that perform best at the task at hand (van der Laan, Polley and Hubbard, 2007). Asymptotically ensemble methods will select the best performing methods for a particular problem from the collection of methods (van der Laan, Polley and Hubbard, 2007). And we show in simulations that in a finite sample ensembles lead to better estimates, as measured by mean square error. Ensembles are also useful because they can estimate more complex functional forms than the underlying methods. And ensembles make estimates more robust—limiting the possibility that a coding error in any one method could lead to invalid conclusions (Dietterich, 2000).

3.1 Constructing the Ensemble via Super learning

Our ensemble estimator will be a weighted average of heterogeneous treatment effect estimators, where the estimators out of sample performance will determine the weights. To construct this weighted average, we use the cross validation based methodology *super learning* introduced in van der Laan, Polley and Hubbard (2007), a method that we will show is closely related to other ensemble methods (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012). Ensembles of methods are increasingly used for diverse problems including supervised text classification (Hillard, Purpura and Wilkerson, 2008) and prediction (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012; van der Laan, Polley and Hubbard, 2007). Both classification and prediction tasks are closely related to the estimation of heterogeneous treatment effects. In classification and prediction, the goal is to estimate a function like $g(k, \boldsymbol{x})$, in order to make an out of sample estimate about a document or future observation. Heterogeneous treatment effects share a similar goal, but take the difference between

response surfaces to estimate the heterogeneous effects. And just as in classification and prediction, identifying features—covariates and treatment assignments—that systematically affect the response surface will improve our estimates of the quantities of interest.

To begin constructing the ensemble, suppose we include M models ($m = 1, 2, \ldots, M$) for estimating heterogeneous effects. For each method m we will define its estimate of $E[Y(k)|T = k] = g_m(k)$ and $E[Y(k)|T = k, \boldsymbol{x}] = g_m(k, \boldsymbol{x})$. Along with the M models, we will suppose that we have a set of weights attached to each of the models $\boldsymbol{w} = (w_1, w_2, \ldots, w_M)$. We will assume that all weights are greater than or equal to zero ($w_m \geq 0$ for all m) and the weights sum to 1 ($\sum_{m=1}^{M} w_m = 1$).

With the weights and models, we define our ensemble estimate of the ATE for conditions k and $k^{'}$ as $\widehat{\phi}(k, k^{'})$,

$$\widehat{\phi}(k, k^{'}) \quad = \quad \sum_{m=1}^{M} w_m g_m(k) - \sum_{m=1}^{M} w_m g_m(k^{'}). \tag{3.2}$$

Analogously, define the ensemble estimate for the CATE for conditions k and $k^{'}$ and covariates $\boldsymbol{x}$ as $\widehat{\phi}(k, k^{'}, \boldsymbol{x})$,

$$\widehat{\phi}(k, k^{'}, \boldsymbol{x}) \quad = \quad \sum_{m=1}^{M} w_m g_m(k, \boldsymbol{x}) - \sum_{m=1}^{M} w_m g_m(k^{'}, \boldsymbol{x}). \tag{3.3}$$

To estimate MCATEs we will take the appropriate averages over CATEs, using a joint distribution to weight the averages. In words, Equations 3.2 and 3.3 show that the ensemble creates a final estimate of a heterogeneous treatment effect by weighting the estimates of the heterogeneous effect from the corresponding models.

The ensembles that we use will weight the predictions from the component models. Following van der Laan, Polley and Hubbard (2007) we determine the weight to attach to each method using the component methods out of sample predictive performance, assessed using cross validation. Out of sample predictive performance (or classification) is used because methods that perform well at individual classification are also likely to perform well at estimating heterogeneous treatment effects. Intuitively, this occurs because a method that

predicts individual out of sample responses well must identify systematic responses to treatments and systematic heterogeneity in response to treatments—exactly the characteristics that lead to accurate estimation of heterogeneous effects. The result is that methods that separate systematic features that assist in prediction also identify systematic differences that represent heterogeneity.[3].

Given this close connection between out of sample prediction and accurate estimation of heterogeneous treatment effects, we create our ensemble in three broad steps (van der Laan, Polley and Hubbard, 2007). In order to weight methods by their out of sample performance, we first generate out of sample predictions for each observation using each of the component methods. To do this, we proceed as if we are performing C-fold cross validation: we randomly divide our data into C subsets ($c = 1, \ldots, C$).[4] For each subset c, we train all M component methods using the participants in the other subsets $\boldsymbol{X}_{-c}$, $\boldsymbol{T}_{-c}$ and $\boldsymbol{Y}_{-c}$. Then, we generate predictions using the trained models and the participants' in subset c's covariates and treatment assignment. The result of this procedure is an $N \times M$ matrix $\widehat{\boldsymbol{Y}}$ where entry $\widehat{Y}_{im}$ contains the out of sample prediction for participant i from method m.

Second, we estimate the weights using the out of sample predictions. For each participant i we regress the true response, Y_i, on the out of sample predictions, $(\widehat{Y}_{i1}, \widehat{Y}_{i2}, \ldots, \widehat{Y}_{im})$. Specifically, we fit the model,

$$Y_i \;=\; \sum_{i=1}^{M} w_m \widehat{Y}_{im} + \epsilon_i \tag{3.4}$$

where ϵ_i is an error term. To ensure that each w_m are weights, we impose two constraints on $\boldsymbol{w}_m$: that the weights sum to 1($\sum_{i=1}^{M} w_m = 1$) and that the weights are greater than

[3]Simple algebra shows the close relationship between prediction and estimation of the response surfaces. To see this, suppose that we use estimator g to predict individual responses for participants assigned to treatment k and with covariates $\boldsymbol{x}$. We can measure its performance with mean square error $E[(Y(k)-g)^2|\boldsymbol{X}]$ which is equal to $var(g|\boldsymbol{x}) + (E[Y(k) - g|\boldsymbol{x}])^2$. It is easy to see that, because we are making the same prediction for all individuals with equivalent k and $\boldsymbol{x}$ that $var(g|\boldsymbol{x}) + (E[Y(k) - g|\boldsymbol{x}])^2 = E[(E[Y(k)|\boldsymbol{x}] - g)^2|\boldsymbol{x}]$, or the mean square error for g in predicting the response surface $E[Y(k)|\boldsymbol{x}]$. It is easy to see that this argument will then carry over if we average over $\boldsymbol{x}$: an estimator with a smaller mean square error with individual prediction will also have a smaller mean square error in estimating $E[Y(k)|\boldsymbol{x}]$

[4]We use 10-fold cross validation below. If there are concerns about sample size, the number of folds can be increased.

zero ($w_m \geq 0$). Fitting this regression is a straightforward quadratic programming problem, whose solution provides a set of weight estimates $\widehat{\boldsymbol{w}}$ that we will use to produce our final ensemble. As we show below, this regression will assign weights to methods based on their predictive ability, but also the method's distinctiveness. Methods that predict well out of sample will receive higher weight. But as van der Laan, Polley and Hubbard (2007) argue, methods that have highly correlated predictions but strong performance will tend to higher predictive weight when aggregated together.

In the third and final step we use the weights and the component methods to generate an ensemble. We fit each of the component models to the entire sample. To create final estimates of interest, we generate synthetic observations (Green and Kern, 2012). For discrete covariates we generate all unique covariate and treatment combinations. For continuous covariates we vary over the range of the covariate. We then use the component methods to generate estimates of the heterogeneous effects for each of the component methods and then use the estimated weights to create a weighted average, as in Equations 3.2 and 3.3. We summarize the steps of the process in Table 1.

Table 1: Three Steps to Generating Ensemble Estimates for Heterogeneous Treatment Effects

1) Generate out of sample predictions for all N observations from all M component methods using C-fold cross validation.
2) Estimate weights for each method based on its out of sample performance using a constrained regression (or other procedure, see Section 3.2).
3) Fit each component model to entire sample and then weight estimates from model using weights estimated in step 2.

3.2 Constructing the Ensemble via Ensemble Bayesian Model Averaging

A closely related ensemble creation procedure is Ensemble Bayesian Model Averaging (EBMA) (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012). EBMA draws on an analogy to Bayesian Model Averaging (BMA) to generate a weighted ensemble to generate predictions. To do this, EBMA utilizes a predictive posterior that is a mixture of component

predictions. Given our focus on dichotomous dependent variables, we note that estimates of $E[Y(k)|\boldsymbol{x}]$, $g(k, \boldsymbol{x})$ are also estimates of $P(Y(k) = 1|\boldsymbol{x})$. In this case, then, we can write out predictive posterior as,

$$\begin{aligned} p(Y(k) = 1|\boldsymbol{x}, \boldsymbol{Y}) &= \sum_{m=1}^{M} \int w_m P(Y(k) = 1|\boldsymbol{x}) p(w_m|\boldsymbol{x}, \boldsymbol{Y}) dw_m \\ &= \sum_{m=1}^{M} \int w_m g_m(k, \boldsymbol{x}) p(w_m|\boldsymbol{x}, \boldsymbol{Y}) dw_m \end{aligned}$$

And if we assume that weights are point masses at the maximum a posterior (MAP) estimate—as is commonly done in the literature (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012)—then this reduces to $p(Y(k) = 1|g_1, g_2, \ldots, g_m, \boldsymbol{x}) = \sum_{m=1}^{M} w_m g_m(k, \boldsymbol{x})$. Our estimate of the CATE for treatment conditions k and $k^{'}$ with covariates $\boldsymbol{x}$ is

$$\widehat{\phi}(k, k^{'}, \boldsymbol{x}) = \sum_{m=1}^{M} w_m g_m(k, \boldsymbol{x}) - \sum_{m=1}^{M} w_m g_m(k^{'}, \boldsymbol{x}). \tag{3.5}$$

This is, of course, equivalent to Equation 3.3, or the formula used to compute our ensemble for estimating heterogeneous treatment effects previously proposed.

Super learning and EBMA share a methodology focused on accurate combinations of component methods. The two methods differ (as presented here) in how the weights are estimated. In Appendix A we provide three ways to estimate the weights for EBMA, including the maximum a posteriori (MAP) methods used in the prior literature (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012) and two ways to obtain the posterior distribution on the weights—Gibbs sampling and a variational approximation (Jordan et al., 1999). While distinct, the methods presented in Appendix A share the same intuition as the regression in Step 2 of the super learner algorithm: the out of sample predictions are used to identify the methods that provide accurate out of sample predictions of individual values.

3.3 Potential Objections to Ensemble Based Methods

Ensemble based methods have demonstrated accurate performance across diverse tasks. Yet, there are potential objections to using ensemble based methods for the estimation of heterogeneous effects. First, we have not provided uncertainty estimates for the quantities of

interest. Across all possible cases we cannot guarantee that estimates of the uncertainty estimates are available. This is because it is unclear how to characterize the uncertainty in estimates from some component methods—such as LASSO (Hastie, Tibshirani and Friedman, 2001) or related methods (Imai and Ratkovic, 2013). But for other methods uncertainty estimates are available and are one of the primary advantages of methods like KRLS (Hainmueller and Hazlett, 2012) and BART (Green and Kern, 2012). When all the component methods have well defined uncertainty estimates it is easy to obtain to uncertainty estimates for the ensemble methods. We can obtain this either through a parametric bootstrap (King, Tomz and Wittenberg, 2000) or we can use simple formula to obtain a closed form for the variance of the ensemble (see Appendix B).

A second potential objection to the use of ensembles is that a cross validation procedure could provide conservative estimates of heterogeneous effects. This could occur because we subset our data to generate out of sample predictions. And analyzing only a subset of data necessarily reduces the statistical power to detect heterogeneous effects. But this objection has limited application to our ensembles: the lack of statistical power only matters if it causes methods that perform poorly in the full sample to receive too much weight in the second stage. This is because our final set of estimates for heterogeneous effects are based on models that use the entire sample, ensuring each of the methods have full power to detect the underlying heterogeneous effects.

4 Monte Carlo Simulations of Ensemble Based Methods

We use a series of monte carlo simulations to show the ensemble accurately estimates heterogeneous effects. The strong performance of the ensemble is not surprising, particularly given ensemble methods strong performance at classification and prediction tasks (van der Laan, Polley and Hubbard, 2007; Hillard, Purpura and Wilkerson, 2008; Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012). Our simulations show that by evaluating methods' predictive performance, ensembles attach greater weight to methods that provide better

estimates of the heterogeneous effects.[5]

We assess the performance of the ensemble across four distinct data generating processes. The simulated data generating processes vary in the number of treatments that have systematic effects, the number of treatments that heterogeneous effects with the included covariates, and the type of covariates that are included in the simulation. In two simulations, *Monte Carlo 1* and *Monte Carlo 2* are *sparse* data generating processes—with many of the simulated treatments specified to have an undetectable systematic effect and only a few specified to have heterogeneous effects. *Monte Carlo 3* and *Monte Carlo 4*, specify data generating processes that are more dense—with more treatments having systematic (and large) effects and heterogeneity across covariates. We provide specific details of each data generation process in Appendix C.

We apply our complete ensemble method to each of the simulated data set. In the monte carlo simulations—and in our applications, below—we form an ensemble with seven methods: (1) LASSO (Hastie, Tibshirani and Friedman, 2001); (2) Elastic-Net, with the mixing parameter set to 0.5 (Hastie, Tibshirani and Friedman, 2001); (3) Elastic-Net, with the mixing parameter set to 0.25 (Hastie, Tibshirani and Friedman, 2001); (4) Find It (Imai and Ratkovic, 2013); (5) Bayesian GLM (Gelman et al., 2008); (6) KRLS (Hainmueller and Hazlett, 2012); and a (7) Support Vector Machine (SVM) (Platt, 1998; Keerthi et al., 2001). In Appendix D we provide details about each methods estimation. After performing 10-fold cross validation to generate out of sample estimates, we use Equation 3.4 to determine each method's weights. After estimating the weight we then apply the models to the entire sample and create our ensemble estimate of the treatment effects implied by the data generating process as a weighted average of the estimates from each of the component methods.

In addition to comparing the performance of our ensemble estimator to the component methods, we will compare its performance to a *naïve* ensemble—where all methods are assumed to contribute equally to the average. We measure the performance of the methods

[5]This contributes to other simulation based evidence on the performance of ensembles on similar tasks (see, for example, van der Laan, Polley and Hubbard (2007)).

using the root mean square error of the estimated heterogeneous treatment effects, with a smaller root mean square error implying more accurate estimates of the heterogeneous effects.[6]

The ensemble method outperforms the other methods across diverse data generating processes in our Monte Carlo simulations. For ease of interpretation, Table 2 presents the performance of each method in terms of the ratio of its root mean squared error to the root mean square error of the ensemble estimate (van der Laan, Polley and Hubbard, 2007). If this is greater than 1, then the ensemble has a smaller root mean square error, or performs better in estimating the heterogeneous effects.

Consider the first two columns in Table 2, showing the results for the sparse data generation processes. Here, we see that methods that assume sparsity perform better—methods such as LASSO and Find It. The ensemble outperforms these component methods, however, and is able to more accurately estimate the heterogeneous effects. The third and fourth columns show that methods that assume a dense set of effects and interactions perform better—such as Elastic Net with $\alpha = 0.25$, KRLS, and Bayesian GLM. And yet, in every instance but one (with Bayesian GLM in Monte Carlo 4) the ensemble outperforms the component methods. Column 5 shows that, as a result, the ensemble estimate has the best average performance across data generating processes. This exemplifies why ensembles are useful: because we never actually know the data generating process, we do not know how well a particular method's assumptions fit the underlying effects (Hastie, Tibshirani and Friedman, 2001). Using ensembles ensures that we use methods that reliably capture the heterogeneous effects for a particular problem.

Figure 1 shows why the ensemble is able to outperform the constituent methods: the methods with a smaller RMSE in out of sample predictions receive more weight. Figure 1 presents the weight attached to the constituent methods (vertical axis) against the method's RMSE. This demonstrates a simple relationship: methods with a smaller RMSE receive

[6] As we detail in Appendix C, we compare all treatment effects relative to a control condition where there is no treatment assigned.

Table 2: Across Diverse Problems, the Ensemble Estimator Performs Best (Mean Square Error of Each Method Relative to Mean Square Error of the Ensemble, Reported)

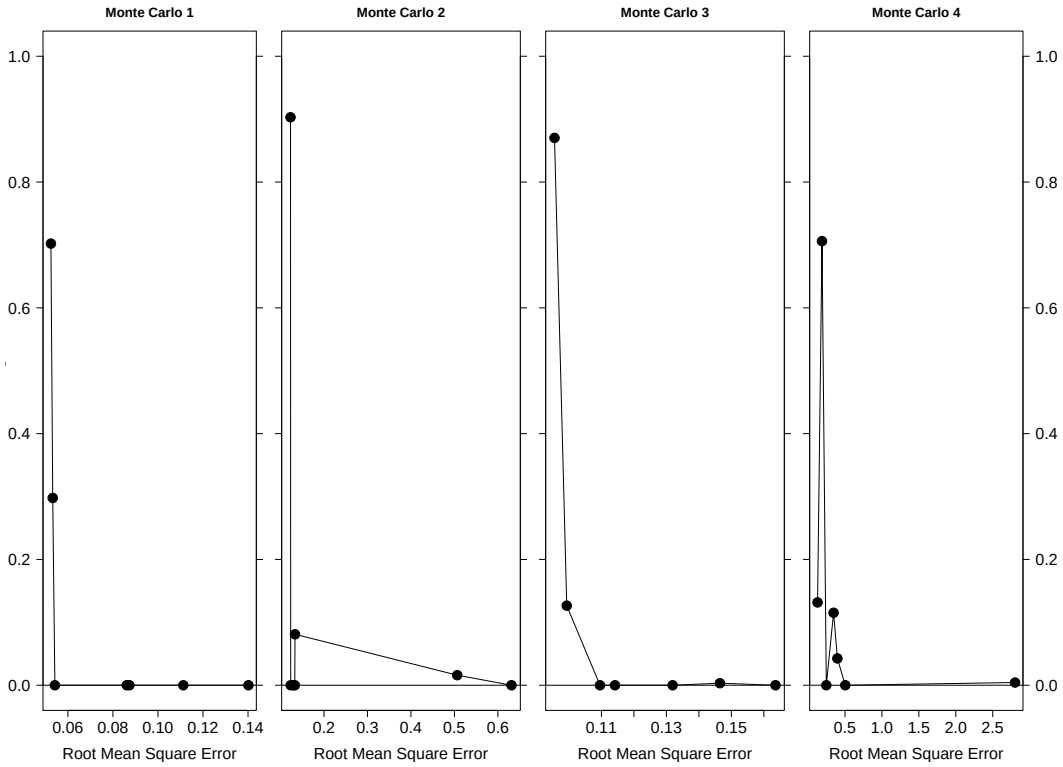
Method	MC 1	MC 2	MC 3	MC 4	Average
LASSO	1.02	1.04	1.40	2.72	1.82
Elastic Net ($\alpha = 0.5$)	1.05	1.05	1.21	1.33	1.20
Elastic Net ($\alpha = 0.25$)	1.69	1.13	1.01	1.02	1.12
Find It	1.03	4.30	1.55	15.12	7.81
Bayesian GLM	2.16	1.13	1.05	0.70	1.05
KRLS	1.67	5.36	1.16	2.14	2.73
SVM-SMO	2.72	1.10	1.73	1.87	1.74
Naïve Average	1.34	1.32	1.07	2.23	1.64

This table shows that the ensemble method outperforms other methods across diverse problems. This table presents the mean square error for each method, divided by the mean square error of the ensemble method. All but one cell entries are greater than 1, indicative of the ensemble's superior performance in estimating heterogeneous treatment effects. The right most column shows that, on average, the ensemble estimator has the lowest mean square error.

greater weight in the final ensemble. When this does not occur exactly—such as Monte Carlo 4— it is because of the estimates from the methods are correlated (van der Laan, Polley and Hubbard, 2007; Montgomery, Hollenbach and Ward, 2012). Aggregated together, the methods with the best performance receive the greatest weight.

The greater weight attached to better performing methods illuminates why ensemble estimates are useful for the estimation of heterogeneous treatment effects and the effects of heterogeneous treatments. Out of sample prediction is closely tied to the estimation of heterogeneous effects. When a method is able to identify systematic variation in the response to treatment, they also implicitly more reliably predict responses out of sample. So weighting methods based on their predictive abilities also implicitly weights methods based on its ability to reliably estimate heterogeneous treatment effects. Together, the simulations show that ensemble methods are able to accurately estimate the effect of heterogeneous treatments. With this strong performance in mind, we turn now to our applications. We use our methods to reveal how constituents reward legislators for securing money in the district

Figure 1: The Ensemble Tends to Place Greater Weight on Methods that More Accurately Measure Heterogeneous Treatment Effects



This figure shows that the ensemble method places more weight on methods that more accurately measure the heterogeneous treatment effect. This occurs even though the method is weighting methods that are performing better at out of sample prediction. This occurs because of the close connection between estimating heterogeneous treatment effects and predicting out of sample.

and how constituents punish legislators for budget deficits.

5 Experiment 1: Rewarded For Type of Expenditure, Not Money

Our first experiment examines the features of credit claiming statements that cause constituents to allocate credit (Grimmer, Messing and Westwood, 2012). Grimmer, Messing and Westwood (2012) argue that legislators' credit claiming statements—a message that legislators use to create the impression they are responsible for some government action—are essential for legislators to receive credit for government spending that occurs in their district (Mayhew, 1974; Grimmer, Messing and Westwood, 2012). Building on the experi-

ments in Grimmer, Messing and Westwood (2012) we conduct a new experiment, to examine how the content of a credit claiming messages affects constituent credit allocation. We assess this information using an experimental template that allows us to vary several features of the message, while maintaining a coherent and realistic message from a legislator.

Participants were assigned either to a control condition (with a 10% chance) or the credit claiming condition (with a 90% chance). Participants in the credit claiming condition read a press release from a fictitious representative who "announced that 17-year old Sara Fisher won 1st place in the annual Congressional art competition". This press release is an example of a common *advertising* press release—a message devoid of policy content intended to increase the legislators' prominence (Mayhew, 1974; Grimmer, 2013). The full text of the condition is in Table 3.

Participants assigned to the credit claiming condition read a message about an expenditure in the district and the fictitious legislator's role in securing that legislation. To assess how different facets of the credit claiming process affects credit allocation we vary five different components of the message: (1) type of expenditure, (2) amount of money, (3) stage in appropriations process, (4) collaboration with other legislators, and (5) representative's political party. Varying the features of message simultaneously allow us to identify the kind of information constituents use when evaluating credit claiming messages, in a way analogous to recent conjoint experiments (Hainmueller and Hopkins, 2013; Hainmueller, Hopkins and Yamamoto, 2013) In Appendix E we summarize the information we provide and how this corresponds to hypotheses about how constituents evaluate legislators' credit claiming efforts. And in Table 3 we summarize the conditions and how the information was provided to participants in our study. We also examine how the effects vary across two relevant respondent characteristics: respondent's *partisan identification*—respondents classify themselves as a *Democrat*, *Republican*, or *Independent/Other*—and *ideological orientation*—respondents classify themselves as *Conservative*, *Liberal*, or *Moderate*.

We examine the effect of legislators' credit claiming efforts on constituents' propensity

Table 3: Examining the Effects of Credit Claiming Statements on Constituent Credit Allocation

Advertising Condition
Headline: Representative (redacted) announces annual Congressional district art competition winner **Body**: Representative (redacted) announced that 17-year old Sara Fischer won 1st place in the annual Congressional district art competition. Sara's winning art, "Medals" was created using colored pencils. Rep. (redacted) said Sara's artwork will be displayed in the U.S. Capitol with other winning entries from districts nationwide.
Credit Claiming Condition
Headline: Representative (redacted) |stageTitle |moneyTitle |typeTitle **Body**: Representative (redacted), |partyMain, |alongMain |stageMain |moneyMain |typeMain. Rep. (redacted) said "This money |stageQuote typeQuote"
|**stageTitle**:[will request/requested/secured] |**moneyTitle**:[$50 thousand/$20 million] |**typeTitle** : [to purchase safety equipment for local firefighters/to purchase safety equipment for local police/to repave local roads, to beautify local parks/for medical equipment at the local planned parenthood/to help build a state of the art gun range] |**partyMain** : [Democrat/Republican] |**alongMain** : [(No text)/and Senator (redacted), a Democrat/ and Senator (redacted), a Republican] |**stageMain** : [will request/requested/secured] |**moneyMain**: [$50 thousand/ $20 million] |**typeMain**: [to purchase safety equipment for local firefighters/to purchase safety equipment for local police/to repave local roads, to beautify local parks/for medical equipment at the local planned parenthood/to help build a state of the art gun range] |**stageQuote** : [would help/would help/will help] |**typeQuote**: [our brave firefighters stay safe as they protect our businesses and homes/our brave police officers stay safe as they protect our property from criminals/keep our roads in safe and working condition, ensuring that our local economy will continue to grow/create parks that add value to the community and provide our children a safe place to play/provide state of the art care for women in our community"/"provide local residents and local, state, and national law enforcement officials a place to sharpen their skills"]
Summary of Conditions **Funding Type**:Planned Parenthood, Parks, Gun Range, Fire Department, Police, Roads **Money**: $ 50 thousand, $20 million **Stage** : Will Requested, Requested, Secured **Who**: Alone, a Senate Democrat, a Senate Republican **Party**: Democrat, Republican

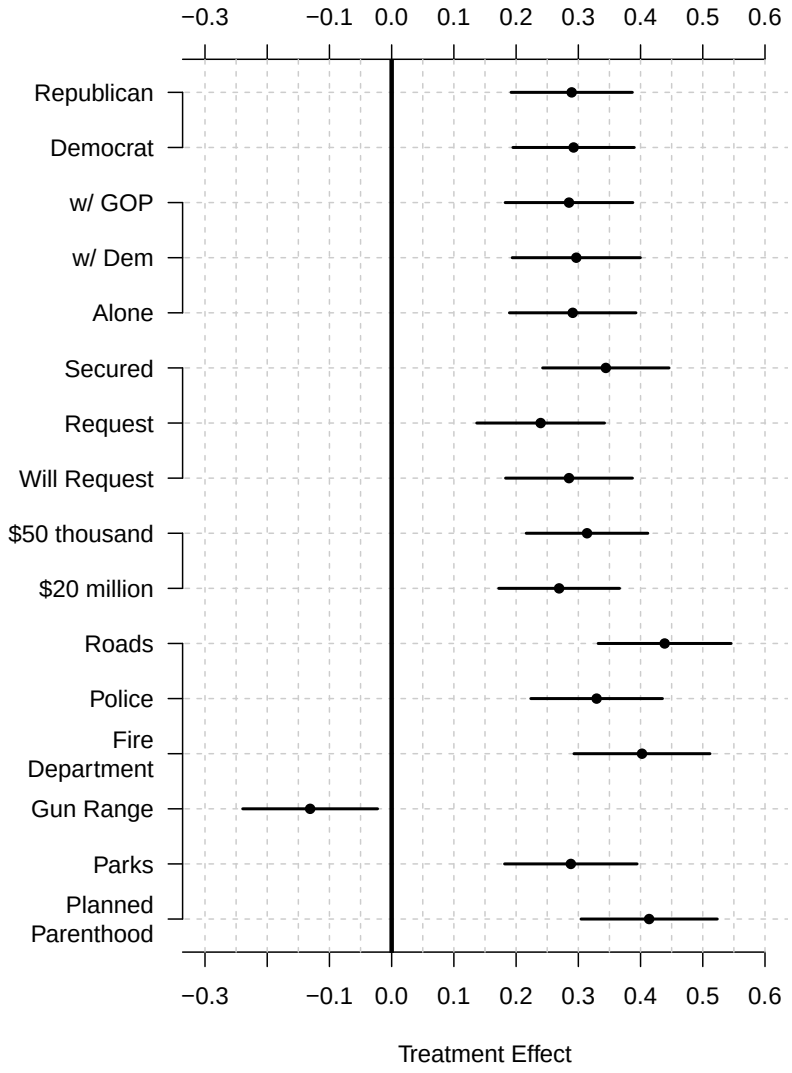
to *Approve* of the representative's performance in office. Specifically, we ask our participants if they "approve or disapprove" of the way the fictitious representative "is performing (his/her) job in Congress". We use the dichotomous response to examine how the content of a legislator's credit claiming messages affects constituent credit allocation.

We recruited 1,074 participants on Amazon.com's Mechanical Turk service, selecting only workers from the United States. We included attention checks at the start and end of our survey to ensure that workers were not satisficing (Berinsky, Huber and Lenz, 2012). After respondents were assigned to a treatment they completed a brief post-survey, that included questions about the legislator and the respondent's own personal political preferences.

Figure 2 shows the marginal average treatment effects, averaging over the accompanying conditions and using the control condition for comparison. The vertical axis shows the information varied in the experiments—including the different types of expenditures, amount of spending, stage in the appropriations process, who announced the expenditure, and the fictitious legislator's party. The point is the marginal average treatment effect of that condition and the lines are 95 percent confidence intervals.

Figure 2 suggests that participants are able to make use of easily available information—such as the type of expenditure—but struggle to include other types of information in their evaluation of credit allocation. The type of expenditure matters considerably—gun ranges decrease approval for representatives (13.1 percentage point decrease, 95 percent confidence interval [-0.25, -0.01]), but other types of projects legislator increase approval over the control condition (37 percentage point increase, 95 percent confidence interval [0.29, 0.46]). Other information, however, has a smaller effect on constituent credit allocation. There is essentially no difference in the credit awarded legislators if they claim credit for $50 thousand or $20 million, who legislators announce with, or whether the legislator is a Democrat or Republican. Securing money does cause an increase in support, relative to stating that the legislator will request or requested the money, with securing causing an 8.1 percentage point great increase in support than requesting or stating an intent to request (95 percent

Figure 2: The Marginal Effects of the Credit Claiming Experiment



This figure shows the marginal effects from the credit claiming experiment. Respondents appear to be evaluating the credit claiming statement based on the type of message, but struggle to include other types of information.

confidence interval [0.02, 0.15]).

The effects in Figure 2 suggest that participants are evaluating legislators' credit claiming efforts by evaluating the type of expenditure, while struggling to use other informa-

tion. If true, then how participants evaluate the type of expenditure will depend on their partisanship and ideology. This is clearest for two of the more polarizing types of expenditures we included: a gun range and funding for planned parenthood. Liberal elites and Democrats tend to vigorously defend planned parenthood, providing cues to like minded citizens that the organization provides valuable services. In contrast, conservatives and Republicans oppose planned parenthood, often working to strip the organization of money (For example, (Kasperowicz, 2013)). Very different cues are available about gun ranges. Many Democrats—particularly liberal-urban Democrats—have argued for increased gun regulation. Republicans and conservatives have argued vigorously for constitutional protection of guns.

To examine how the message content affects credit allocation across participants with different partisan identifications and ideological orientation we use our ensemble method.[7] We apply the ensemble method using 10-fold cross validation, then use Equation 3.4 to estimate the weights attached to each method. The first column of Table 4 shows the weights attached to each method for the ensemble used to generate the effects for this experiment. Three methods receive non-zero weight: LASSO (0.62), KRLS (0.24), and Find It (0.14). In generating the final effects, we compare all treatment effects to the control advertising condition.[8]

Figure 3 shows how the effect of the *type* of project claimed depends on the participant's partisan and ideological identification. On the right-hand side vertical axis we vary the type of expenditure announced and within each type of expenditure we vary participants ideological orientation and partisan identification. To ease interpretation, we draw lines to connect the heterogeneous responses to the same type of expenditure.

Figure 3 reveals substantial heterogeneity in the effect of the *type* of project on con-

[7]Together our experiment has $6 \times 2 \times 3 \times 2 \times 3 + 1$ conditions—a *very* heterogeneous treatment, along with the 9 unique partisan and ideological characteristics. Given our sample size limitations, we examine only pairwise interactions between our treatments in our method—reducing the number of potential conditions from 217 to 98 total conditions.

[8]Because we include methods that do not have easily obtained uncertainty estimates, we are unable to provide uncertainty estimates.

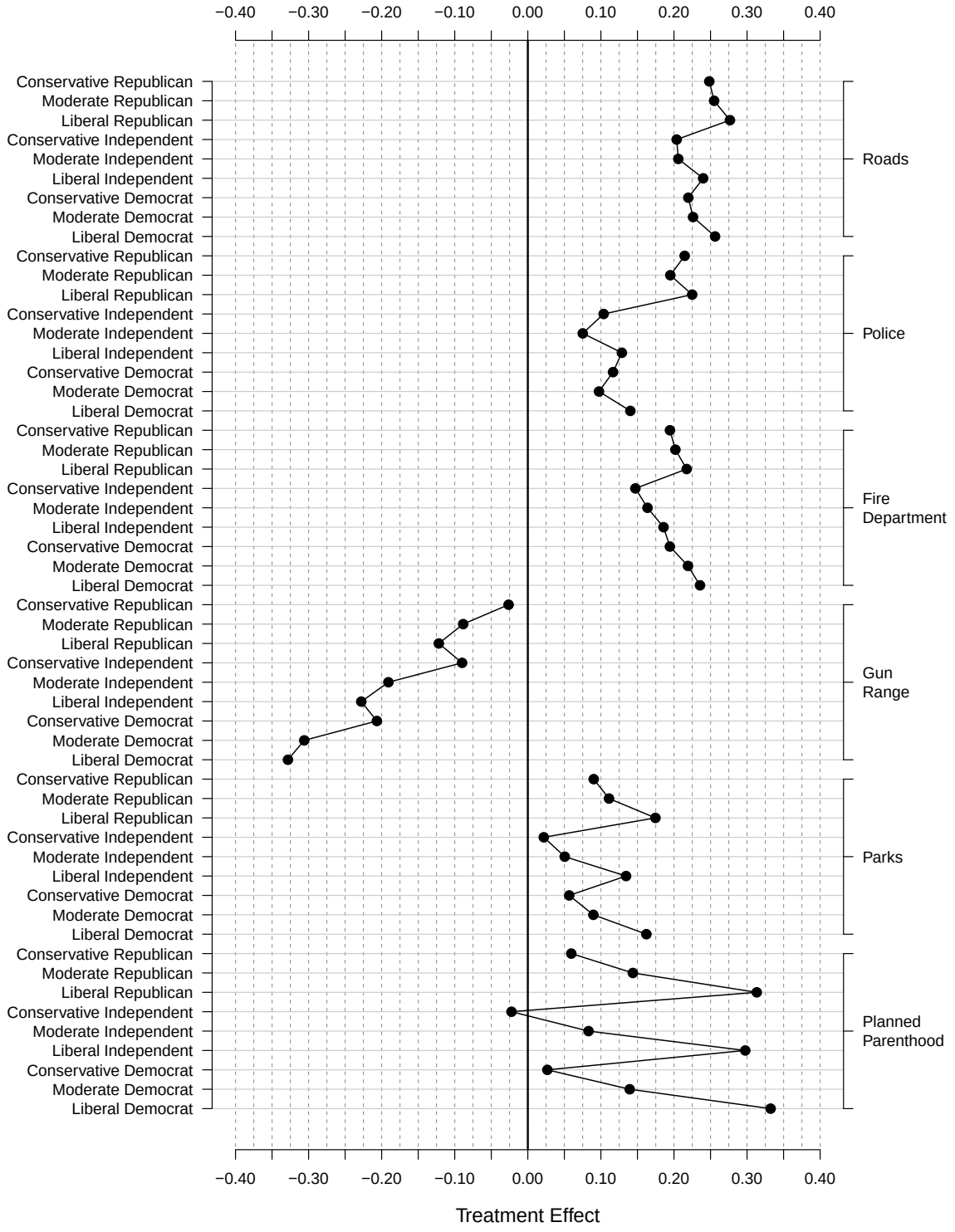
Table 4: The Weight Attached to Methods Varies Across Experiments

Method	Credit Allocation	Deficit Punishment
LASSO	0.62	0.00
Elastic Net (α =0.5)	0.00	0.00
Elastic Net (α = 0.25)	0.00	0.00
Find It	0.14	0.10
Bayesian GLM	0.00	0.00
KRLS	0.24	0.81
SVM-SMO	0.00	0.09

stituent credit allocation, consistent with constituents evaluating the type of expenditure when evaluating legislators' credit claiming statements. Consider the response to money directed to planned parenthood. Liberal respondents, regardless of partisan affiliation, substantially boost support for the representative who claims credit for securing money for planned parenthood. The fictitious legislator claiming credit for funds for planned parenthood caused a 30 percentage point increase in approval rating for liberals. Conservatives, however, have a more muted—and even negative—response to legislators who claim credit for planned parenthood spending. Indeed, claiming credit for money delivered to planned parenthood causes a decrease in approval ratings among independent conservatives and only a small increase among other conservatives.

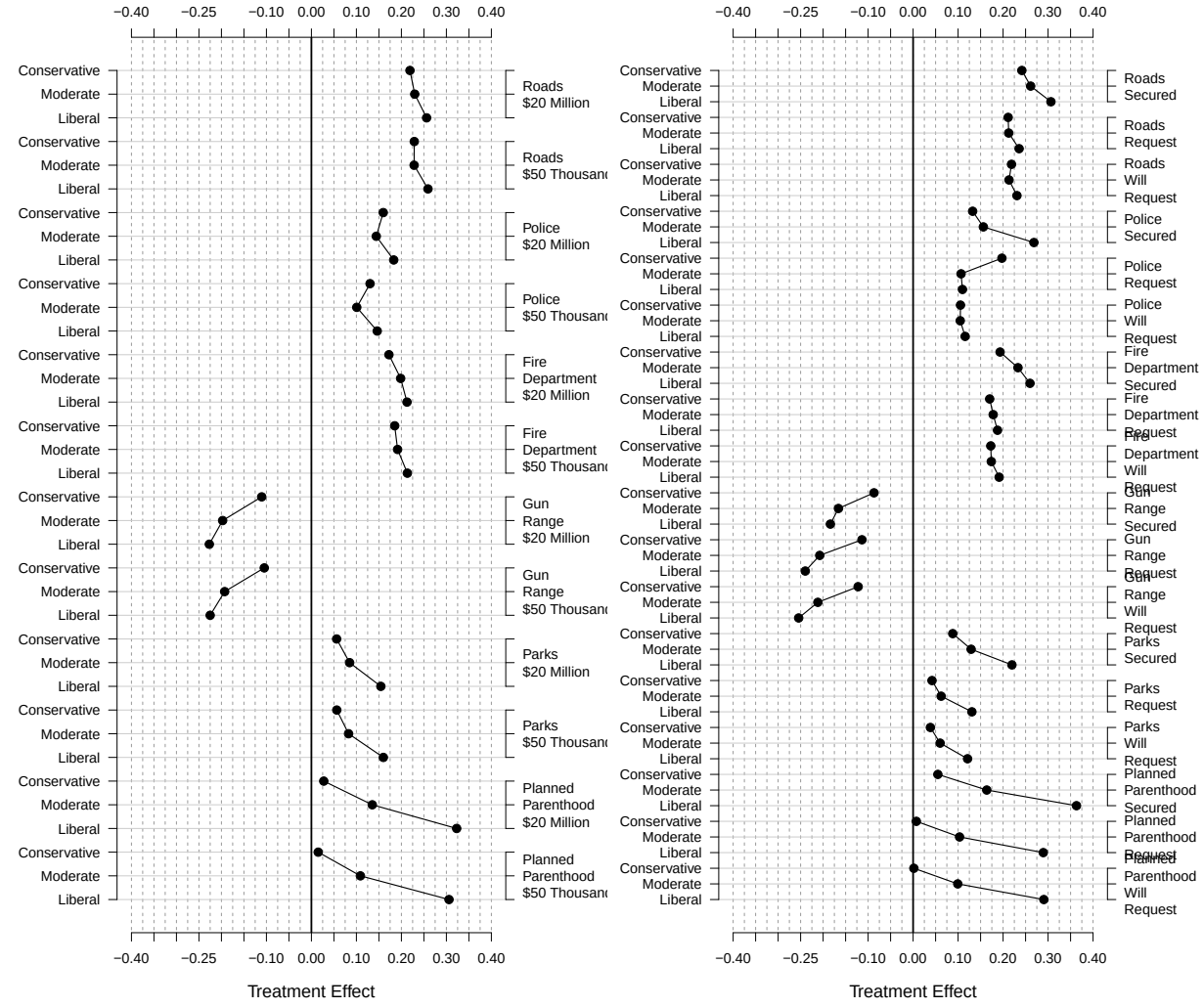
The effect of claiming credit for gun ranges also varies across participants. Constituents who are likely to have a negative attitude about guns have a strong negative reaction to legislators claiming credit for a new gun range. Claiming credit for a gun range causes an over 30 percentage point decline in a legislators' approval rating among liberal Democrats. But constituents who are likely to have a less negative view of guns have a much more muted response to the gun range: claiming credit for a gun range causes only a 5 percentage point decline among conservative Republicans. Conservatives and Republicans do not, however, simply have a smaller magnitude to all credit claiming efforts. Claiming credit for delivering money to police causes a 21 percentage point increase in approval among Republicans, while only an 11 percentage point increase among Democrats and independents.

Figure 3: Constituent Partisanship and Ideology Predicts Differences in the Effectiveness of Credit Allocation



This figure shows substantial heterogeneity in how participants respond to the different types of money legislators claim credit for securing.

Figure 4: The Money Secured or the Stage in the Process Appears to be Less Important



This figure shows that constituents struggle to include information about the amount of money legislators claim credit for securing (left-hand plot) or the stage in the process that the money is in (right-hand plot).

The heterogeneity in response to the type of credit claiming message is consistent with constituents evaluating the type of expenditure when allocating credit for spending in the district. Figure 4 shows that constituents tend to be less responsive to other pieces of information in the credit claiming statements. The left-hand plot in Figure 4 shows how the effect of the type of project and the amount of money (right-hand vertical axis) for the project varies across constituent ideology.[9] This figure shows that constituents—across types of expenditure and ideological orientation— are largely unaffected by the amount claimed. Indeed the parallel lines are indicative of constituents who are largely unable to incorporate information about the size of the expenditure into their overall evaluations. Consistent with evidence from Grimmer, Messing and Westwood (2012), the size of the expenditure appears to matter little when constituents are allocating credit to legislators. The right-hand plot in Figure 4 shows how the effect of credit claiming messages varies across type of money secured and stage in the appropriations process (right-hand vertical axis) and constituents' ideological orientation (left-hand vertical axis). While there is evidence that constituents do reward the fictitious legislator slightly more for *securing* rather than requesting money, the increase from having secured an expenditure is much smaller than the heterogeneity across the type of expenditure.

This section uses the ensemble method for estimating heterogeneous treatment effects to show that constituents tend to evaluate easily available information—such as the type of money secured—but struggle to include other types of information, such as the amount secured or the stage in the appropriations process. Together, this shows how our heterogeneous treatment effect estimator reveals substantively interesting variation in the effects of treatment.[10]

[9] In the Supplemental Appendix we present both plots with both ideology and partisanship. We do not present them here for space and presentation concerns, but the plots are evidence for the same point: participants are less responsive to the amount of money secured and the stage in the process for the expenditure.

[10] In the Supplemental Appendix, we present the heterogeneous treatment effect estimates for the other treatment conditions.

5.1 Experiment 2: Punishment for Budget Deficits

Legislators claim credit for spending to cultivate a personal vote with constituents. Yet, in recent years a growing movement of conservatives, the *Tea Party*, has criticized expenditures as wasteful (Skocpol and Williamson, 2011). Grimmer, Westwood and Messing (2013) design an experiment to assess how the criticism of government spending affects how constituents allocate credit. Grimmer, Westwood and Messing (2013) show that labeling an expenditure as contributing to the budget deficit undermines the credit legislators' receive. We apply our ensemble method to assess how the effect of the criticism varies across constituent characteristics—revealing that the effect of budget criticism varies substantially, depending on constituents' ideological orientation.

Grimmer, Westwood, and Messing's (2013) experiment couples a legislator claiming credit for an expenditure with criticism about the budget implications of the spending. For realism about the magnitude of the effects, this experiment utilizes participants' actual representatives, rather than fictitious representatives used in the first study. Table 5 contains the content of the experiment's three conditions. In the *credit claiming* condition participants are presented with their house member claiming credit for an $84 million highway expenditure in their Congressional district. To customize the paragraph about each participant's legislator, the representative's name is inserted at |representativeName and the participant's state at |state in the text.

Two *budget criticism* conditions vary the source of the information about how the spending will affect the federal deficit. The *CBO Budget Information* condition includes this same credit claiming about a highway expenditure, but pairs it with information about the budget consequences of the expenditure from the non-partisan Congressional Budget Office (CBO). This condition has an official statement from the CBO, which includes the overall cost of the program and that expenditure would be deficit spending. In the *Partisan Information* condition, participants receive budget information from a political figure likely to criticize the participant's member of Congress: the opposing party's national chairperson. Partici-

pants with a Democratic representative see a statement from Reince Priebus—chair of the Republican national committee—and participants with a Republican member of Congress see a statement from Debbie Wasserman-Schultz—chair of the Democratic national committee. Both the Democratic and Republican national committees regularly criticize opposing partisans for actions in Congress—ensuring that our treatments are ecologically valid and replicate the kind of criticism that occurred in response to the stimulus. To add further ecological validity, we make the opposing party chairpersons more critical of the expenditure. In addition to the CBO information on the budget consequences of the bill, the opposing party chairpersons also assert that "the spending bill is wasteful."

As in Experiment 1, we examine the effects of the message on whether participants approve of their representative, again examining whether participants approve or disapprove of their representative. This experiment was administered to 1,166 participants, using a census matched US sample from a Survey Sampling International (SSI) panel. Participants were assigned to conditions, the treatments were administered and then a post-survey was administered. Grimmer, Westwood and Messing (2013) shows that the budget criticism have a strong and negative effect on legislators' approval ratings. The budget information from the CBO causes an overall decrease of 8.2 percentage points (95 percent confidence interval [-0.16, -0.01]) and the partisan information causes a similar overall decrease in approval of 7.7 percentage points (95 percent confidence interval [-0.15, -0.00]).

To examine how the effects of the intervention vary across constituent and legislator characteristics, we apply our ensemble method to the experiment. Table 4 shows the diverse methods that receive weight for this experiment: including KRLS (0.81), Find It (0.10), and SVM (0.09). We then use the ensemble to compute marginal conditional average treatment effects for combinations of respondent characteristics and particular treatments, with all effect sizes based on a comparison to the credit claiming message without criticism.

Following Imai and Strauss (2011) and Imai and Ratkovic (2013), we first use the ensemble method to assess the constituents who have the most negative and positive response to

Table 5: Content Across Conditions, Experiment 2

Headline: Representative |representativeLastName Announces $84 Million for Local Road Projects
Body: Representative |representativeName (|party - |state) announced that the Department of Transportation Federal Highway Administration has released $84 million for local road and highway projects. Representative |representativeName said 'I am pleased to announce that we will receive $84 Million from the Federal Highway Administration. It is critical that we support our infrastructure to ensure that our roads are safe for travelers and the efficient flow of commerce.' This funding will add lanes to |state highways.
CBO Budget Information: The nonpartisan Congressional Budget Office reported that the spending bill is wasteful and contributes to the growing federal deficit. "This bill contributes to federal spending without identifying a new source of revenue or off-setting budget cuts. Accounting for the total cost of this program across all Congressional districts, the bill costs taxpayers $36.5 billion, all of which is added to the deficit and compounded with interest."
Partisan Information: [Debbie Wasserman-Schultz, Chair of the Democratic National Committee/Reince Preibus, Chair of the Republican National Committee] said that the spending bill is wasteful and contributes to the growing federal deficit. "This bill contributes to federal spending without identifying a new source of revenue or off-setting budget cuts. Accounting for the total cost of this program across all Congressional districts, the bill costs taxpayers $36.5 billion, all of which will be added to the deficit and compounded over time with interest.
Key
|representativeName: Representative's name
|party: Representative's party
|state: Represenative's state

the criticism. Table 6 show who has the most negative and positive response to the budget criticisms for Democratic and Republican representatives. The most negative responses for both Democratic and Republican representatives come from constituents who identify as conservative—strong conservatives for Republican representatives, conservatives for Democratic representatives. For Democratic legislators the most positive response comes from strong liberals who are also strong Democrats—the most positive response for Republican representatives is from moderate Republicans.

Table 6: Characteristics of Constituents with the Most Negative Response to Budget Criticism for Democrat and Republican Representatives

Democrat Representatives, Most Negative Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
-0.36	High School	\$50k-\$80k	50-63	Female	White	Cons.	Dem.
-0.36	High School	\$50k-\$80k	64 +	Female	White	Cons.	Dem.
-0.36	Grad. Degree	\$50k-\$80k	50-63	Female	White	Cons.	Dem.
Democrat Representatives, Most Positive Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
0.28	College	< \$50k	36-49	Male	White	Strong Lib.	Strong Dem.
0.27	College	< \$50k	36-49	Female	White	Strong Lib.	Strong Dem.
0.27	College	< \$50k	< 37	Male	White	Strong Lib.	Independent
Republican Representatives, Most Negative Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
-0.32	College	< \$50k	50-63	Male	Non-white	Strong Cons.	Strong Rep.
-0.32	College	\$50k-\$80k	50-63	Male	White	Strong Cons.	Strong Rep.
-0.31	College	< \$50k	50-63	Male	White	Strong Cons.	Strong Rep.
Republican Representatives, Most Positive Response							
Effect	Education	Income	Age	Gender	Race	Ideology	Party
0.25	College	< \$50k	< 37	Male	White	Moderate	Republican
0.24	College	< \$50k	< 37	Male	White	Moderate	Republican
0.24	College	< \$50k	< 37	Female	White	Moderate	Republican

Table 6 reveals who has the strongest response to the budget criticism treatment, suggesting that conservatives are particularly likely to punish legislators for deficit spending and that strong liberals are particularly likely to reject the criticism. To examine how the effect of the criticism varies across different types of respondents, Figure 5 shows how the effect of the budget criticism affects credit allocation for Democratic and Republican representatives (right-hand axis) and for constituents with varying ideological orientations (strong liberal, liberal, moderate, conservative, strong conservative) and partisan affiliations (strong Democrat, Democrat, Independent, Republican, strong Republican). This figure shows that strong liberals—regardless of partisan affiliation or (almost) regardless of partisan alignment with the representative —tend to have a *positive* response to the budget criticism. This aligns well with cues from political elites, who have suggested that deficit spending does less harm than members of the political right emphasize. Further, moderate Republicans appear unaf-

fected by information about spending's budget implications. Conservatives, however, have a particularly negative response to learning about the budget implications of an expenditure. And as Table 6 illuminates, strong conservatives have a particularly negative response to the criticism.

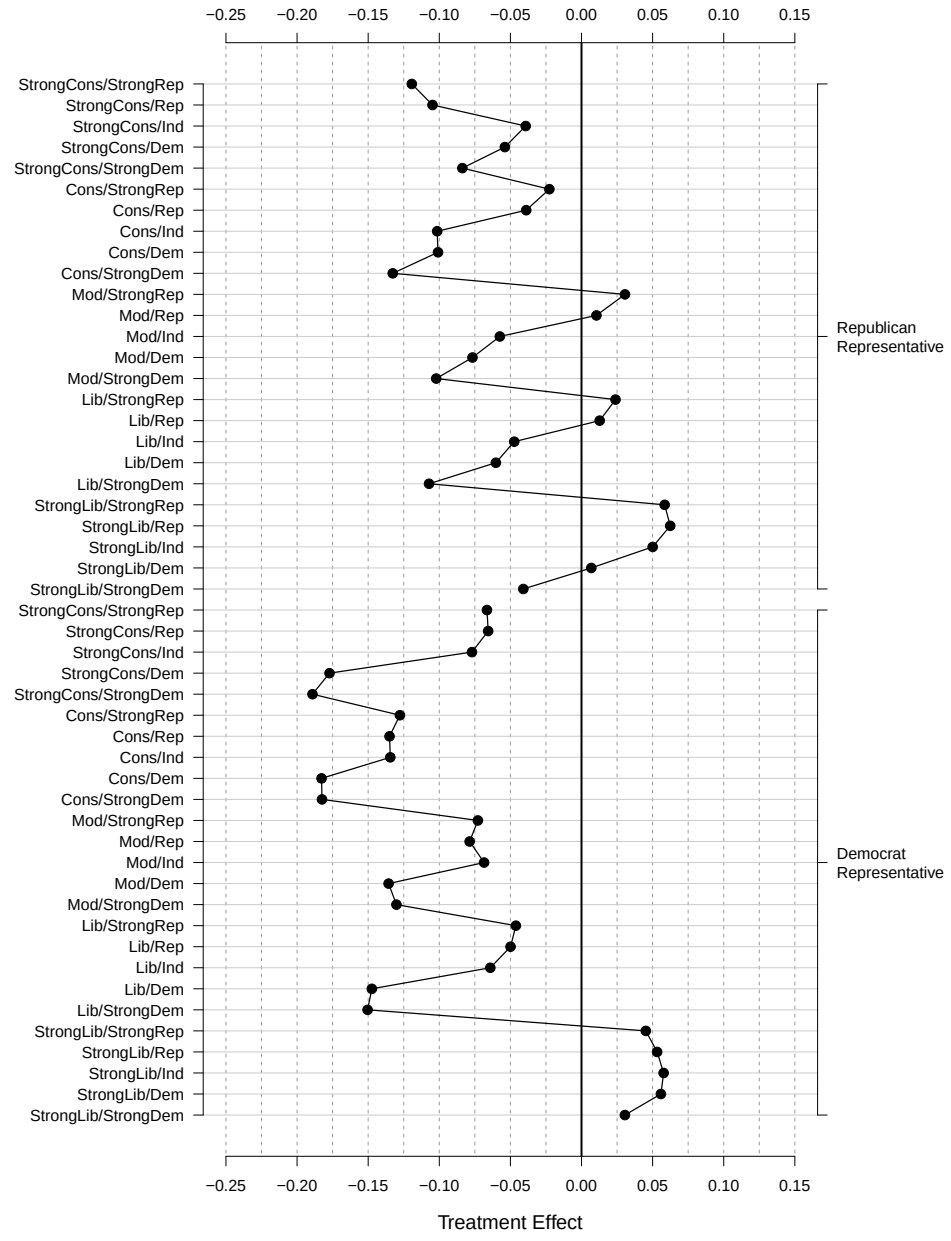
The variation in response to the budget criticism reveals how Tea Party rhetoric affects how constituents respond to budget criticism. For participants likely to be receptive to the rhetoric—conservatives—the budget criticism causes sharp punishment of the elected official. But for other constituents, the budget criticism is ineffective. Strong liberals–who have likely learned of the arguments for deficit spending—are unperturbed by the criticism and continue allocating credit to legislators for spending—in some cases greater expenditures. Together, the heterogeneity in response illuminates how critical rhetoric can make it costly for Republicans to claim credit for spending, because they depend on the support of strong conservatives in the primary. Democrats, however, do not have the same costs, because the most ideological members of their base—strong liberals, have a positive response to the budget criticism.

6 Conclusion

We have shown how weighted ensembles of methods provide reliable estimates of heterogeneous treatment effects. Across diverse problems the ensemble is able to provide accurate estimates, because it attaches more weight to methods that provide more accurate estimates for the particular task. Then applying the weighted ensemble to two experiments, we show how respondents evaluate easy to obtain information when accessing a legislators' credit claiming statement and the substantial ideological variation in how participants punish, or reward, legislators for deficit spending.

We have shown how ensembles provide accurate and reliable estimates of heterogeneous treatment effects across diverse problems. Ensembles do not, however, obviate the need for developing new methods for estimating heterogeneous effects. Quite to the contrary, the ensembles only work *because of the impressive innovations of the constituent methods*. New

Figure 5: Strong Liberals are Unresponsive to Budget Criticisms



This figure shows how the response to budget criticism depends on constituents' characteristics (left-hand vertical axis). For both Democrats and Republicans (right-hand vertical axis), we see that Strong liberals are unresponsive to the budget criticism, or the criticism may cause an increase in legislators' approval ratings. Strong conservatives, however, a much more negative reaction to the credit claiming efforts. This is consistent with the rise of the Tea Party movement.

innovations in the estimation of heterogeneous effects, then, will improve the performance of the ensemble estimates.

Ensembles do suggest, however, a new way to evaluate methods for heterogeneous effects: the weight attached to a method. The weight attached to a method is a useful method because a proposed method for estimating heterogeneous treatment effects will be useful if it is both accurate—reliably estimating heterogeneous effects—and novel—providing estimates that are difficult to obtain otherwise (van der Laan, Polley and Hubbard, 2007). And methods that are both novel and accurate receive more weight in the ensemble. This validation shows the power of recent innovations in the estimation of heterogeneous effects: both KRLS and Find It receive substantial weight when applied to actual experimental data.

Far from replacing individual methods, then, ensemble estimates of heterogeneous effects provide a way to make use of the impressive new innovations in the estimation of heterogeneous effects. By pooling together the methods, we make the most of new methods and the new experimental data.

A Estimating Weights for EBMA

In this appendix we describe the posterior distribution for EBMA and provide three ways to estimate the weights. Following prior literature (Raftery et al., 2005; Montgomery, Hollenbach and Ward, 2012) we assume that our predictive posterior is a mixture of the component methods. We will suppose that the weights are drawn from a uniform distribution (or a Dirichlet($\mathbf{1}$)). We will suppose that each observation i is drawn from one of the M component models. Denote the model with a $M \times 1$ indicator vector $\boldsymbol{\tau}_i$ where $\tau_{im} = 1$ when observation i is drawn from model m and all other entries are zero. We will suppose that $\boldsymbol{\tau}_i \sim \text{Multinomial}(\boldsymbol{w})$. Finally, given a realization of $\boldsymbol{\tau}_i$ with $\tau_{im} = 1$ we will suppose that the out of sample prediction for observation i assigned to treatment k $Y(k)_i$ is drawn from a Bernoulli distribution, with chance of success $\pi = g_{im}(k, \boldsymbol{x})$ or $\widehat{Y}_{im}(k)$ in the notation above.

Together this implies the following model

$$\begin{aligned} \boldsymbol{w} &\sim \text{Dirichlet}(\mathbf{1}) \\ \boldsymbol{\tau} &\sim \text{Multinomial}(\boldsymbol{w}) \\ Y_i(k)|\tau_{im} = 1, \boldsymbol{x} &\sim \text{Bernoulli}\left(\widehat{Y}_{im}(k)\right) \end{aligned}$$

and the following posterior distribution for the weights,

$$p(\boldsymbol{w}, \boldsymbol{\tau}|\widehat{\boldsymbol{Y}}, \boldsymbol{x}, \boldsymbol{Y}) \propto \prod_{i=1}^{N}\prod_{m=1}^{M}\left[w_m \times \left(\widehat{Y}_{im}(k)^{Y_i(k)} \times (1-\widehat{Y}_{im}(k))^{1-Y_i(k)}\right)\right]^{\tau_{im}}$$

We provide three ways to estimate weights with this posterior: an Expectation-Maximization (EM) algorithm, a Gibbs sampler, and a variational approximation. Each derivation is straightforward and available in previous work on estimation in mixture models.

A.1 EM Algorithm

The EM algorithm proceeds in two steps. To begin, initialize estimates for the weights w_m^t where t will index the iteration. Then, we compute the E-step. For each observation i and each model m compute $\widehat{\tau}_{im}$ which is equal to

$$\widehat{\tau}_{im}^t = \frac{w_m^t\left[\widehat{Y}_{im}(k)^{Y_i(k)} \times (1-\widehat{Y}_{im}(k))^{1-Y_i(k)}\right]}{\sum_{l=1}^{M} w_l^t\left[\widehat{Y}_{im}(k)^{Y_i(k)} \times (1-\widehat{Y}_{im}(k))^{1-Y_i(k)}\right]}$$

Computing the M step is straightforward, with the new estimates of the weight for model m, w_m^{t+1} given by

$$w_m^{t+1} \propto 1 + \sum_{i=1}^{N}\widehat{\tau}_{im}^t$$

Estimation of the EM-algorithm proceeds until the change in the parameters (or other summary of changes) drops below a predetermined threshold. The EM estimates,

A.2 Gibbs Sampler

A Gibbs sampler provides estimates of the posterior. This facilitates estimation of the uncertainty in the weights when calcuating ATEs and CATEs. Like the EM algorithm, the

steps of the Gibbs sampler are well established. Again, initialize weights w_m^t where t tracks the iteration of the sampler. We then sample in two stages. First, we draw $\widehat{\boldsymbol{\tau}}_i^t$,

$$\widehat{\boldsymbol{\tau}}_i^t \sim \text{Multinomial}(1, \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iM})$ and

$$\theta_{im}^t = \frac{w_m^t \left[\widehat{Y}_{im}(k)^{Y_i(k)} \times (1 - \widehat{Y}_{im}(k))^{1-Y_i(k)}\right]}{\sum_{l=1}^M w_l^t \left[\widehat{Y}_{im}(k)^{Y_i(k)} \times (1 - \widehat{Y}_{im}(k))^{1-Y_i(k)}\right]}$$

Conditional on the drawn indicator vectors, $\widehat{\boldsymbol{\tau}}_i$, we draw the weights, $\boldsymbol{w}^t$,

$$\boldsymbol{w}^{t+1} \sim \text{Dirichlet}(\boldsymbol{\eta})$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_M)$ and

$$\eta_m = 1 + \sum_{i=1}^N \widehat{\tau}_{im}^t.$$

After a burn in period and convergence is diagnosed, the sampler is run to approximate the posterior distribution of weights. These weights can then be used to estimate the ATEs and CATEs.

A.3 Variational Approximation

A third method for estimating the posterior on the weights is with a variational approximation, a deterministic method for approximating the full posterior. Variational approximations make a simplifying assumption about the posterior and then finds the member of this simpler, though still general, functional family that provides the closets approximation to the full posterior, as measured by the Kullback-Leibler divergence. We will approximate the posterior distribution for $\boldsymbol{w}$ and $\boldsymbol{\tau}$ with the simpler functional form $q(\boldsymbol{w}, \boldsymbol{t}) = q(\boldsymbol{w})q(\boldsymbol{\tau})$. By the independence assumptions in our data, this implies that we can write the approximating function as $q(\boldsymbol{w})q(\boldsymbol{\tau}) = q(\boldsymbol{w}) \prod_{i=1}^N q(\boldsymbol{\tau}_i)$.

Standard arguments for variational approximations of exponential family distributions (see Jordan et al. (1999); Bishop (2006)) leads to the form of the posterior approximations

and the update steps. A standard derivation shows that $q(\boldsymbol{\tau}_i)$ is a Multinomial distribution, with parameter $\boldsymbol{\theta}_i$ where

$$\theta_{im} \quad \propto \quad \exp\left(\mathrm{E}[\log w_m] + \log\left[\widehat{Y}_{im}(k)^{Y_i(k)} \times (1-\widehat{Y}_{im}(k))^{1-Y_i(k)}\right]\right)$$

where $\mathrm{E}[\log w_m]$ is taken over the approximating distribution and dependent on $q(\boldsymbol{w})$. A second standard calculation shows that $q(\boldsymbol{w})$ is a Dirichlet($\boldsymbol{\eta}$) distribution with η_m equal to

,

$$\eta_m = 1 + \sum_{i=1}^{N} \theta_{im}$$

This implies that $\mathrm{E}[\log w_m] = \psi(\eta_m) - \psi\left(\sum_{l=1}^{M} \eta_l\right)$ where $\psi(\cdot)$ is the digamma function. After initializing values of $\boldsymbol{\eta}^t$ the formulas are applied iteratively to update the parameters until the change in the parameters (or change in a lower bound) drops below a sufficient level for convergence. The approximating posterior distribution on the weights with the converged parameter estimates can then be used to reflect posterior uncertainty in the weights.

B Uncertainty Estimates

We do not present uncertainty estimates for our ensemble because we make use of methods whose uncertainty estimates are still open questions (Hastie, Tibshirani and Friedman, 2001; Imai and Ratkovic, 2013). But if an ensemble is composed of methods that have uncertainty estimates, then it is straightforward to compute uncertainty estimates. One potential approach is a parametric bootstrap method, if it is possible to perform a parametric bootstrap for the component methods. If this is possible, uncertainty calculations are straightforward. First, for each of the M component methods and for treatment conditions k and $k^{'}$ and covariates $\boldsymbol{x}$, use the parametric bootstrap to draw T realizations ($t = 1, 2, \ldots, T$) for each of the M methods. Then, we for each realization t we compute the weighted average of the methods and then take the difference, to obtain estimate $\widehat{\phi}(k, k^{'}, \boldsymbol{x})^t$. Uncertainty estimates can then be obtained from this vector of simulations.

We can also obtain uncertainty estimates with a closed form expression. First, note that

$$\begin{aligned}\text{var}\left(\widehat{\phi}(k,k^{'},\boldsymbol{x})\right) &= \text{var}\left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x})\right) + \text{var}\left(\sum_{m=1}^{M} w_m g_m(k^{'},\boldsymbol{x})\right) \\ &\quad -2\text{cov}\left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x}), \sum_{m=1}^{M} w_m g_m(k^{'},\boldsymbol{x})\right).\end{aligned}$$

Estimating var $\left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x})\right)$ is straightforward. By the law of total variance,

$$\text{var}\left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x})\right) = \sum_{m=1}^{M} w_m \text{var}\left(g_m(k,\boldsymbol{x})\right) + \sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x})^2 - \left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x})\right)^2$$

We can then estimate cov $\left(\sum_{m=1}^{M} w_m g_m(k,\boldsymbol{x}), \sum_{m=1}^{M} w_m g_m(k^{'},\boldsymbol{x})\right)$ with further assumptions (such as constant covariance across $\boldsymbol{x}$) or other restrictions.

C Details on Monte Carlo Simulations

We specify four data generating processes for our Monte Carlo simulations. Each of the Monte Carlo simulations build off the simulations in Imai and Ratkovic (2013).

Monte Carlo 1 For this simulation we have a sparse data generating process with discrete covariates. Specifically, we suppose that for all 2500 observations that,

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi\left(\boldsymbol{\beta}\boldsymbol{T}_i + \gamma \boldsymbol{X}_i + \sum_{k=1}^{46}\sum_{j=1}^{2} \eta_{jk} X_{ij} \times T_{ik}\right) \end{aligned} \tag{C.1}$$

where:

- Φ is the standard Normal CDF
- $\boldsymbol{T}_i$ is a 46-element treatment indicator vector. Suppose that $\boldsymbol{p}$ is a 47 element long vector equal to $(\frac{1}{47}, \frac{1}{47}, \ldots, \frac{1}{47})$. Then we draw $T_i \sim$ Multinomial($\boldsymbol{p}$) and if all elements of $\boldsymbol{T}_i$ are equal to zero then this corresponds with a control condition.
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{46})$ are coefficients for T_i. We set $\beta_1 = 2$, $beta_2 = 1, \beta_3 = 0.5, \beta_4 = -1, \beta_5 - 2$. For k from 6 to 46 we draw $\beta_k \sim$ Uniform$(-0.07, 0.07)$.

- $\boldsymbol{X}_i$ is a 2-element long vector of covariates, with $X_{i1} \sim$ Bernoulli0.4 and $X_{i2} \sim$ Bernoulli(0.6) .

- $\boldsymbol{\eta}$ is a vector of interaction terms for each treatment and covariate. We suppose that the first five treatments have systematic interactions with the covariates. The remaining eta values are assumed to be drawn from a Uniform$(-0.1, 0.1)$ distribution.

We then assess the RMSE by generating all possible treatment and covariate combinations and comparing to the actual estimated effects.

Monte Carlo 2 For this simulation we maintain the same basic structure as in Monte Carlo 1, but change the discrete covariates to continuous covariates. Specifically, we suppose that in Equation C.2 that for each i we generate $a_i \sim \text{Normal}(0, 1)$, $b_i \sim \text{Normal}(0, 1)$, and $c_i \sim \text{Normal}(0, 1)$. We then compute,

- $X_{i1} = \sin(a_i) \times b_i + \cos(c_i) * a_i$

- $X_{i2} = \exp\left(\frac{a_i}{10}\right) \times (b_i^2 + \sin(c_i))$

Because the continuous covariates don't allow us to exactly estimate the treatment effects for every possible valuable, we vary across a range of each variable to compare the actual and estimate treatment effects.

Monte Carlo 3 Monte Carlo 3 provides a *dense* data generating process, with many more treatments having a systematic and large effect—and many more having hetergeneous treatment effects. We suppose again the basic structure

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi\left(\boldsymbol{\beta}\boldsymbol{T}_i + \gamma \boldsymbol{X}_i + \sum_{k=1}^{46}\sum_{j=1}^{2} \eta_{jk} X_{ij} \times T_{ik}\right) \end{aligned} \tag{C.2}$$

where:

- Φ is the standard Normal CDF

- $\boldsymbol{T}_i$ is a 46-element treatment indicator vector. Suppose that $\boldsymbol{p}$ is a 47 element long vector equal to $(\frac{1}{47}, \frac{1}{47}, \ldots, \frac{1}{47})$. Then we draw $T_i \sim$ Multinomial($\boldsymbol{p}$) and if all elements of $\boldsymbol{T}_i$ are equal to zero then this corresponds with a control condition.

- But now we suppose that many more of the treatments have systematic effects. Specifically we suppose for each k ($k = 1, \ldots, 46$) that we draw $n_k \sim$ Bernoulli(0.5). And then we draw the coefficients,

$$\beta_k \sim \begin{cases} \text{Normal } (-1, 0.1) \text{ If } n_k = 1 \\ \text{Normal } (1, 0.1) \text{ If } n_k = 0 \end{cases}$$

- And we suppose that there are interactions between covariates and the treatments for all the covariate and treatment pairs. We suppose each for each j and k we draw $n_{jk} \sim$ Bernoulli(0.5). And then for each η_{jk} we draw,

$$\eta_{ij} \sim \begin{cases} \text{Uniform}(-1, -0.5) \text{ If } n_{jk} = 1 \\ \text{Uniform}(0.5, 1) \text{ If } n_{jk} = 0 \end{cases}$$

Monte Carlo 4 This Monte Carlo simulation generates the covariates as in Monte Carlo 2 and coefficients as in Monte Carlo 3.

D Details of Ensemble Creation

We apply seven methods to estimate the heterogeneous treatment effects.

1) LASSO: We estimate the LASSO using the `glmnet` (Friedman, Hastie and Tibshirani, 2010). We use cross validation to determine the penalty parameter, using mean square error, and the binomial family. We predict values with the penalty parameter that minimizes the mean square error.

2) Elastic-Net $\alpha = 0.5$: We estimate the elastic net using the `glmnet` (Friedman, Hastie and Tibshirani, 2010). We use cross validation to determine the penalty parameter, using mean square error, and the binomial family. We predict values with the penalty parameter that minimizes the mean square error.

3) Elastic-Net $\alpha = 0.25$: We estimate the elastic net using the `glmnet` (Friedman, Hastie and Tibshirani, 2010). We use cross validation to determine the penalty parameter, using mean square error, and the binomial family. We predict values with the penalty parameter that minimizes the mean square error.

4) Bayesian GLM: We use the logit link in the binomial family in the `arm` package (Gelman and Hill, 2007)

5) Find It: we use the `FindIt` package (Imai and Ratkovic, 2013). We search for the lambda parameters and use the glmnet option.

6) KRLS: we use the `KRLS` package, using a gaussian kernel.

7) SVM: we use the `RWeka` and `rJava` packages using the SMO command, with the polynomial kernels.

E Details on Experiment 1

In this section we provide additional details on the theoretical reasons for each of the five components of our template.

Type : We vary the *type* of expenditure directed to the district, allowing us to assess how differences in preferences about the kinds of expenditures affect the credit allocated legislators (Lazarus and Reiley, 2010). We include commonly announced types of expenditures—such as police grants, fire department grants, and roads. We also include more controversial types of funding—such as planned parenthood and funds for the gun range, because they generate clear hypotheses about how the treatment response will vary with respondent characteristics.

Money : We also vary the *amount* of money secured, providing a second assessment of how constituents respond to the dollar amount of a project (rather than other features of

the project) (Grimmer, Messing and Westwood, 2012).

Stage : Legislators often announce projects throughout the appropriations process. To assess how constituents allocate credit at different stages, we vary the *stage* in the appropriations process of the expenditure.

Collaboration : Legislators also often announce expenditures with other members of Congress (Grimmer, 2013) and there are diverse theories about how this partnership affects the credit legislators receive (Shepsle et al., 2009; Chen, 2010). We vary who legislators *announce with* to test how the partnership affects the credit legislators receive.

Partisanship Finally we vary the *partisanship* of the representative, because partisan cues may affect how constituents evaluate messages from legislators.

References

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20:351–368.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. Springer.

Chen, Jowei. 2010. "Electoral Geography's Effect on Pork Barreling in Legislatures." *American Journal of Political Science* 54(2):301–322.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 41(1):266–298.

Dietterich, Thomas. 2000. "Ensemble Methods in Machine Learning." *Multiple Classifier Systems* pp. 1–15.

Friedman, Jerome, Trevor Hastie and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1):1.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2(4):1360–1383.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, Andrew, Jennifer Hill and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5(1).

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiment: Design, Analysis, and Interpretation*. W.W. Norton & Company.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grimmer, Justin. 2013. *Representational Style: What Legislators Say and Why It Matters*. Cambridge University Press.

Grimmer, Justin, Sean J. Westwood and Solomon Messing. 2013. "The Impression of Influence: How Words and Money Cultivate a Personal Vote." Stanford University Mimeo. Book Manuscript.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2012. "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation." *American Political Science Review* 106.

Hainmueller, Jens and Chad Hazlett. 2012. "Kernel Regularized Least Squares: Moving Beyond Linearity and Additivity Without Sacrificing Interpretability." Massachusetts Institute of Technology, Mimeo.

Hainmueller, Jens, Daniel Hopkins and Teppei Yamamoto. 2013. "Causal Inference in Conjoint Analysis: Understanding Multi-Dimensional Choices via Stated Preference Experiments." MIT Mimeo.

Hainmueller, Jens and Daniel J. Hopkins. 2013. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Toward Immigrants." MIT Mimeo.

Hartman, Erin, Richard Grieve, Roland Ramshai and Jasjeet S. Sekhon. 2012. "From SATE to PATT: Combining Experimental with Observational Studies." University of California, Berkeley Mimeo.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer.

Hillard, Dustin, Stephen Purpura and John Wilkerson. 2008. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research." *Journal of Information Technology & Politics* 4(4):31–46.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Jordan, Michael et al. 1999. "An Introduction to Variational Methods for Graphical Models." *Machine Learning* 37:183–233.

Kasperowicz, Pete. 2013. "GOP Seeks Planned Parenthood Study with Hope to Strip Funding." Politico.com.

Keerthi, S.S., S.K. Shevade, C. Bhattacharyya and K.R.K. Murthy. 2001. "Improvements to Platt's SMO Algorithm for SVM Classifier Design." *Neural Computation* 13(3):637–649.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347–361.

Lazarus, Jeffrey and Shauna Reiley. 2010. "The Electoral Benefits of Distributive Spending." *Political Research Quarterly* 63(2):343–355.

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

Montgomery, Jacob M., Florian M. Hollenbach and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20(3):271–291.

Platt, J. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*, ed. B. Schoelkopf, C. Burges and A. Smola. MIT Press.
URL: *http://research.microsoft.com/ jplatt/smo.html*

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review* 133:1155–1174.

Shepsle, Kenneth A. et al. 2009. "The Senate Electoral Cycle and Bicameral Appropriations Politics." *American Journal of Political Science* 53(2):343–359.

Skocpol, Theda and Vanessa Williamson. 2011. *The Tea Party and the Remaking of Republican Conservatism*. Oxford University Press.

van der Laan, Mark, Eric Polley and Alan Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).