Political Analysis Advance Access published April 2, 2011

doi:10.1093/pan/mpr005

Estimating Treatment Effects in the Presence of Noncompliance and Nonresponse: The Generalized Endogenous Treatment Model

Kevin M. Esterling
Department of Political Science, UC–Riverside, 900 University Ave., Riverside, CA 92506
e-mail: kevin.esterling@ucr.edu (corresponding author)

Michael A. Neblo
Department of Political Science, Ohio State University, 2114 Derby Hall, 154 N Oval Mall, Columbus, OH 43210
e-mail:neblo.1@osu.edu

David M. J. Lazer
Departments of Political Science and Computer Science, Northeastern University, 301 Meserve Hall, Boston, MA 02115
e-mail: d.lazer@neu.edu

If ignored, noncompliance with a treatment or nonresponse on outcome measures can bias estimates of treatment effects in a randomized experiment. To identify and estimate causal treatment effects in the case where compliance and response depend on unobservables, we propose the parametric generalized endogenous treatment (GET) model. GET incorporates behavioral responses within an experiment to measure each subject's latent compliance type and identifies causal effects via principal stratification. Using simulation methods and an application to field experimental data, we show GET has a dramatically lower mean squared error for treatment effect estimates than existing approaches to principal stratification that impute, rather than measure, compliance type. In addition, we show that GET allows one to relax and test the instrumental variable exclusion restriction assumption, to test for the presence of treatment effect heterogeneity across a range of compliance types, and to test for treatment ignorability when treatment and control samples are balanced on observable covariates.

1 Introduction

In a causal analysis, the treatment effect is defined as a comparison between a subject's outcome if administered the treatment and the same subject's outcome under the control, or the difference between the subject's potential outcomes in each state (Rubin 1974; Holland 1986). Since the same subject cannot be observed in both states, the causal comparison is a counterfactual and thus necessarily depends on missing data. In an ideal randomized experiment, one can estimate the average treatment effect (ATE) across subjects through a simple comparison of treatment and control group averages; the control group can supply the missing potential outcomes for those in the treatment group and vice versa. In many experimental settings, however, the researcher can only encourage, not compel, subjects to participate in the experimental tasks, and treatment noncompliance and nonresponse on the outcomes can destroy

Authors' note: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Alberto Abadie, Janet Box-Steffensmeier, Bear Braumoeller, Kosuke Imai, Luke Keele, Gary King, William Minozzi, Jas Sekhon, Craig Volden, and two anonymous referees for very helpful comments. Earlier versions of this paper were presented at the Kennedy School of Government, Harvard University, May 2007; the Annual Meetings of the American Political Science Association, Chicago, IL, August 2007; the West Coast Experiments Conference, Riverside, CA, May 2008; and at the Economic Science Association 2008 Annual Meeting, Pasadena, CA, June 2008. For replication data and code, see Esterling, Neblo, and Lazer (2011).

a randomization. In this situation, the researcher must adjust for these compliance processes within the analysis in order to make a correct counterfactual comparison between treatment and control groups.[1]

If one could observe the compliance process for each subject, this adjustment to identify the treatment effect would be straightforward—apples in the treatment group could be compared to apples in the control group, and oranges to oranges.[2] In many situations, however, the compliance process is driven by unobserved variables, where it is difficult to describe or measure subjects' ability or motivation to comply with an assigned treatment. For a given experiment, some subjects are of the ''type'' to comply, others are not, and often types in this sense are latent and unobserved. For example, the application below examines the effects from participating in a deliberative experiment, but to date there is no available measure that captures the propensity to deliberate.

We propose a new method, the generalized endogenous treatment (GET) model, to identify the treatment effect counterfactual in the case where noncompliance and nonresponse are conditioned on unobservables. GET exploits behavioral measures that often are generated within an experiment in order to measure a subject's latent *compliance type*, and it simultaneously conditions on the measured compliance variable as a means to correct the counterfactual comparison and to identify the treatment effect.

In this paper, we implement the GET model using Bayesian estimation methods (e.g., Imbens 1997; Horiuchi, Imai, and Taniguchi 2007) and demonstrate its properties using simulation methods and an application from a deliberative field experiment.

2 Motivation

GET is a generalization of the method of instrumental variables (IV). IV is widely used to address the problem of noncompliance with the treatment when the compliance process is unobservable (Angrist, Imbens, and Rubin 1996). IV remains inconsistent in the presence of outcome nonresponse, however, as compliers tend to have a different probability of response than noncompliers, and each may have a different distribution of potential outcomes (Barnard et al. 2003, 302; Horiuchi et al. 2007, 674). To address this limitation, Frangakis and Rubin (1999) proposed the method of principal stratification (see also Frangakis and Rubin 2002; Barnard et al. 2003; Mealli et al. 2004; Horiuchi et al. 2007). The Frangakis and Rubin (FR) approach to principal stratification identifies treatment effects by assuming that the outcome is independent of both treatment compliance and missing responses within strata of a categorical compliance type variable, a conditional independence assumption they label ''latent ignorability.'' These categories include a set of ''compliers'' who take up the treatment if and only if they are assigned to the treatment. The set of ''noncompliers'' is the complement to compliers, which potentially includes ''never takers'' who do not take up the treatment whether or not they are assigned to the treatment group, and ''always takers'' who take up the treatment whether or not they are assigned to the treatment.[3]

The key insight of FR is that compliance type is a pretreatment covariate and hence cannot be affected by treatment assignment or exposure to the treatment. Under latent ignorability, conditioning on compliance type makes the treatment assignment, treatment exposure, and missing responses independent of the outcome (exposure to the treatment and outcome response are ignorable), and the causal effect of a treatment is identified within strata of the compliance variable (Frangakis and Rubin 1999; Barnard et al. 2003). As a consequence, treatment effects are identified even when latent dependencies exist among compliance, nonresponse, and the outcomes of interest, that is, even in the case of compliance processes driven by unobservables.

In the FR approach, however, principal stratification retains the significant limitation that it must treat compliance type as missing data for many subjects in the sample. For example, compliers and never takers cannot be distinguished among those assigned to the control group since these never takers are not

[1] One can construct nonparametric bounds for treatment effects without assumptions about the compliance process (e.g., Imai and Yamamoto 2010), but such bounds are often too wide to convey much useful information, and often researchers wish to produce point estimate distributions.

[2] Nonparametric matching can construct appropriate treatment and control comparison groups when the compliance and response processes can be captured fully using observed variables (Abadie et al. 2001; Imbens 2004; Rosenbaum and Rubin 1985).

[3] One other compliance category logically exists, the ''defiers'' who take up the treatment if and only if they are not assigned to the treatment, but researchers typically assume this category is never observed under a monotonicity assumption.

given the opportunity to refuse the treatment. In this approach, missing compliance data may be imputed (Barnard et al. 2003; Horiuchi et al. 2007), provided a further assumption is met, the ''compound exclusion restriction'' that requires assignment has no direct effect on potential outcomes or response. An estimation procedure that treats subjects' compliance type as missing data ignores any information, however, on compliance behavior that is potentially available through the study. Exploiting such information can eliminate any need for the exclusion restriction, an assumption that is often implausible in practice, and can improve the efficiency by which latent compliance type is estimated.

To incorporate such additional information on compliance type that may be available for all subjects, we propose the GET model. Building on the random effect approach in Aakvik, Heckman, and Vytlacil (2005), the GET model extends principal stratification to the case where multiple indicators of compliance exist for all subjects, by estimating each subject's latent compliance type in a measurement model (e.g., Trier and Jackman 2008). In contrast to the FR approach to principal stratification, GET measures compliance based on observed behavior instead of a model-based imputation that relies heavily on the explanatory power of covariates.

Since GET measures the compliance type both for those assigned to the active treatment and those assigned to the control, it affords four major improvements over FR's approach to principal stratification:

1. Because it incorporates additional information regarding compliance type, GET has the potential to improve the efficiency of treatment effect estimates compared to FR. By construction, the latent compliance type is a powerful predictor of compliance with the treatment. Conditioning on this latent characteristic helps to reduce uncertainty in the imputation of the compliance type indicators, and so less uncertainty is passed on to treatment effect estimates.
2. Because compliance type is measured for all subjects, GET does not require any form the ''compound exclusion restriction'' required in ordinary principal stratification (see, e.g., Mealli and Rubin 2003). The exclusion restriction requires the randomized assignment to have no direct effect on nonresponse or the outcomes. Principal stratification ordinarily requires the exclusion restriction as an identifying constraint when compliance type is missing for the controls (Mealli et al. 2004), although Bayesian parametric approaches (e.g., Hirano, Imbens, and Zhou 2000) can relax this restriction when using informed priors. The GET model does not require the exclusion restriction in any form, and the existence of any effects of treatment assignment can be modeled and tested.
3. Because it is based on a measurement model, each subject's compliance type is a continuous variable rather than a partially observed categorical variable. FR's method takes subjects' compliance behavior in the treatment condition as a categorical measure of compliance type. As a consequence, treatment effects themselves can only be all-or-none; compliers can experience a treatment effect but noncompliers cannot. With GET, treatment effect heterogeneity can be made a function of the (continuous) compliance measure.
4. The latent compliance variable enters the equations modeling outcomes and compliance as a common random effect. As a result, one can retrieve the correlation among compliance and outcomes induced by the latent compliance type after conditioning on covariates. This correlation is a useful test for ignorability after balancing on observed covariates using a matching procedure.

3 The GET Model

GET is a generalization of the random effect models proposed in Aakvik et al. (2005), Miranda and Rabe-Hesketh (2006), and Terza (1998). It extends the random effect approach to identify treatment effects found in Aakvik et al. (2005) to account for nonresponse, linking this literature to the concept of principal stratification.

3.1 *Assumptions for Identifying Causal Effects*

Define the experiment **e** as a study that randomly assigns subjects to treatment and control conditions in an effort intended to estimate the ATE on a set of outcomes. In field experiments, a treatment condition often is an encouragement for subjects to take up some treatment, rather than an application of the treatment itself. For a given subject, cells in this experiment are defined by a partitioned data vector $\mathbf{e} = (\mathbf{z}, \mathbf{o}, \mathbf{c}, T,$

$\mathbf{o_{Pt}}$, $\mathbf{x}$) with elements of $\mathbf{e}$ defined below (individual indexes are suppressed). To identify treatment causal effects in GET, across subsamples or across the entire sample, we rely on several assumptions.[4]

We first assume *randomization* of the treatment assignment;[5] the exogenous vector $\mathbf{z}$ indicates subjects' randomized assignment to a treatment condition. Randomization ensures that there exists subjects of the full range of compliance types in both the treatment and control assignment conditions. Next we assume *stability* in the treatment (SUTVA); the treatment itself is binary and there is no interference among units. Under stability, $\mathbf{z}$ contains only a scalar Z that is one if assigned to the treatment, zero otherwise. Stability reduces the number of treatment levels and so greatly simplifies the modeling problem.

There are K separate (but potentially dependent) experimental outcomes in $\mathbf{o}$. In principle, we wish to compare potential outcomes from the experiment, $\Delta_k = E(O_{1k} - O_{0k})$, in expectation, where O_{1k} is defined as the kth potential outcome that would be observed if the subject were assigned to the treatment and O_{0k} is the potential outcome that would be observed if the same subject was assigned to the control. For each subject, we only observe the outcome corresponding to her realized treatment assignment, $O_k \equiv ZO_{1k} + (1 - Z)O_{0k}$. Define the treatment outcome vector $\mathbf{o}$ of length K to be the potential outcomes that are observed for a given subject. If the subject did not respond on the outcome measurement, her outcomes are missing. For example, in the application below, the treatment outcome vector is observed if and only if the subject responded to a follow-up survey administered at the completion of the experiment.

To simplify the modeling task, we assume (very) *strong monotonicity*, that the treatment is only available to those assigned to the treatment, $Z = 1$, and that some of those who are encouraged to take the treatment actually take it. This form of monotonicity is ensured by a research design that prevents those in the control from gaining access to the treatment, a design common in clinical studies (Mealli et al. 2004, 210), and by randomization of a large number of subjects.[6]

The vector $\mathbf{e}$ contains a subvector $\mathbf{c}$ of length M of endogenous variables that indicate separate, but potentially dependent, self-selected compliance choices. Included in $\mathbf{c}$ is an indicator variable C_t that equals 1 if the subject took up the treatment, and 0 otherwise; C_t is missing only among those assigned to the control ($Z = 0$). An indicator variable C_r equals one if the subject responded on the outcome measures, zero otherwise. In addition, $\mathbf{c}$ also includes any other indicator variables, $\mathbf{c_{\neg tr}}$, measuring compliance type. For example, in the application below, subjects chose whether or not to respond to multiple waves of a survey (as in Horiuchi et al. 2007). For additional elements of $\mathbf{c_{\neg tr}}$, one could also build in separate compliance tasks into the experimental design or one could use pretreatment covariates that predict compliance such as survey items or the subjects's past history of compliance that may be available through a survey firm's panel records. To evaluate the causal effect of the treatment, define a treatment exposure indicator T that equals 1 if the subject received the treatment and 0 otherwise.

The vector $\mathbf{e}$ also may contain pretreatment variables that can reduce variability within the model or that are also known to determine potential outcomes.[7] For example, the subvector $\mathbf{o_{pt}}$ of length $\boldsymbol{K}$ may contain the pretreatment values of the outcomes corresponding to those measured in $\mathbf{o}$, although these pretreatment variables are not required. Since there typically are omitted variables that codetermine the corresponding elements of $\mathbf{o_{pt}}$ and $\mathbf{o}$, we take $\mathbf{o_{pt}}$ as an endogenous vector of covariates (Morgan and Winship 2007, 71). The subvector $\mathbf{x}$ contains exogenous covariates that may be related to the outcomes; $\mathbf{x}$ does not include a constant.

Finally we state the *compound exclusion restriction* from IV analysis, which requires that the assignment itself has no effect on the potential outcomes other than through the treatment as well as no effect on the probability of response. This restriction enables the analyst to identify causal effects without direct

[4] If an additional assumption of random sample selection (Imai et al. 2008) is not met, then only sample treatment effects are identified.

[5] This assumption is not strictly required since it is possible to achieve balance on unobservables by other assignment mechanisms, but we do not consider alternative mechanisms here.

[6] As implemented below, GET can distinguish compliers from never takers but not from always takers. As a result, our implementation can only work where subjects in the control do not have access to the treatment or where always takers are a very small portion of the sample (as in Barnard et al. 2003, 303). If always takers do exist in the sample, one would need to collect indicators that can distinguish them from compliers, such as building in a placebo task that they are not encouraged to take, and then measure compliance in a categorical model.

[7] Note that such pretreatment covariates can be helpful for GET but are not required, as with matching.

knowledge of every subject's compliance type (and again, usually those in the treatment condition reveal their type while those in the control do not). This assumption is easily violated since the encouragement itself can cause subjects to change their behavior (Hirano, Imbens, and Zhou 2000; Mealli and Rubin 2003), such as when a subject who is offered the treatment but refuses has a reduced probability of responding to a post-treatment survey.[8] GET does not require the exclusion restriction since it measures compliance type for all subjects.

3.2 *Parametric Model*

To be consistent with the application below, assume the outcomes are dichotomous, $O_k \in \{0, 1\}$, and the data generating processes (DGPs) are Bernoulli. The model is easily extended to any DGP within the class of linear exponential functions (Skrondal and Rabe-Hesketh 2004). We take the realization of O_k to be a function of a latent index variable O_k^*. We wish to estimate a row vector $\boldsymbol{\theta}_\mathbf{k} = (\boldsymbol{\alpha}_\mathbf{k}, \boldsymbol{\beta}_\mathbf{k}, \boldsymbol{\lambda}_k)$ of structural parameters in each of K regressions:

$$O_k^* = \alpha_{0k} + \alpha_{1k} T + \alpha_{2k} O_{pt(k)} + \mathbf{x}\boldsymbol{\beta}_\mathbf{k}^{'} + \lambda_k \eta_1 + \boldsymbol{\epsilon}_\kappa \tag{1a}$$

$$O_k = \begin{cases} 1 & \text{if } O_k^* > 0, \\ 0 & \text{if } O_k^* \leqslant 0. \end{cases} \tag{1b}$$

where η_1 is an estimated measure of each subject's propensity to comply with the experiment, to be defined below.[9] We state the distributional assumptions for η_1 and $\boldsymbol{\epsilon}_k$ after the measurement model below.

As in all models based on principal stratification, GET identifies the structural parameters by conditioning on subjects' latent compliance type. For simplicity, we assume that the experimental design prevents subjects assigned to the control from having access to the treatment. In the FR approach to principal stratification, under this design the sample can contain only two types of subjects: compliers who would take up the treatment if asked and never takers who do not take up the treatment even when asked. In contrast, GET takes compliance type as a continuous and unidimensional latent variable. We measure compliance type, η_1, using the behavioral indicators $\mathbf{c}$ and M additional regressions of the form

$$C_m^* = \alpha_{0m} + \mathbf{x}\boldsymbol{\beta}_\mathbf{m}^{'} + \lambda_m \eta_1 + \boldsymbol{\epsilon}_m \tag{2a}$$

$$C_m = \begin{cases} 1 & \text{if } C_m^* > 0, \\ 0 & \text{if } C_m^* \leqslant 0. \end{cases} \tag{2b}$$

with $m \in \{t, r, \neg\text{tr}\}$, again giving distributional assumptions for η_1 and $\boldsymbol{\epsilon}_m$ below.

Using equation (2a), GET uses the behavioral indicators $\mathbf{c}$ to estimate a factor η_1 measuring subjects' latent tendency to comply with all aspects of the experiment, along with a vector of factor coefficients, $\boldsymbol{\lambda}_\mathbf{m}$. That is, η_1 is a measure of each subject's compliance type. This measurement model shares the properties of item response models (Patz and Junker 1999; Trier and Jackman 2008). Since item response models estimate a full distribution for the latent trait for each individual in the sample, the measured variable η_1 accounts for the uncertainty in the estimate of each subject's compliance type.

For the distributions of $(\eta_1, \boldsymbol{\epsilon}_k, \boldsymbol{\epsilon}_m)$, we impose the standard assumptions in dichotomous random effect models required for identification (Skrondal and Rabe-Hesketh 2004). We assume $\eta_1 \sim N(0,1)$; $\boldsymbol{\epsilon}_m \sim N(0,1)$, and $\text{cov}(\eta_1, \boldsymbol{\epsilon}_m)$ for all m; $\boldsymbol{\epsilon}_k \sim N(0,1)$, $\text{cov}(\eta_1, \boldsymbol{\epsilon}_k) = 0$ for all k; and $\text{cov}(\boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_k) = 0$ for all m, k.[10] In addition, one of the free parameters in $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_\mathbf{m}, \boldsymbol{\lambda}_\mathbf{k}\}$ must be set to 1 to scale the latent variable η_1. Finally,

[8] The exclusion restriction is most plausible in a double-blinded experiment, but blinding is impossible in encouragement designs that dominate field experimental methods.

[9] Note that this is the same specification as the outcome equation in Horiuchi et al. (2007), which specifies the linear index as $\alpha_h Z C_t + \beta_h (1 - Z) C_t$. One can see this by substituting $\alpha_h - \beta_h$ for α_{1k}, β_h for λ_k, and the dichotomous compliance variable C_t for η_1 in equation (1a) and rearranging.

[10] We use the normal distribution to structure all latent variables and error terms. This assumption is not required and can easily be relaxed using alternate distributions or nonparametric latent variable methods.

we assume that η_1 is orthogonal to the $\mathbf{x}$ vector; if an element of $\mathbf{x}$ is thought to fail this assumption, it can be treated as endogenous by allowing it to load on η_1.

The GET approach assumes that subjects' unobserved compliance type can be characterized by a continuous latent variable, η_1 (see Aakvik et al. 2005). GET simultaneously includes the common factor η_1 as a latent control variable in the K outcome equation (1a), with coefficient vector $\boldsymbol{\lambda}_\mathbf{k}$. By including η_1, the GET model holds subjects' behavioral ''compliance type'' (measured as η_1) constant as a way to identify the structural parameters. Here, we use a *linearity* assumption that requires that the outcome and response choice variables depend on an index that is linear in parameters. This assumption accommodates linear, convex, or concave functional relationships between the compliance variable and the linear indexes for each choice. With the assumption of linearity, strong monotonicity, latent ignorability, and the exclusion restriction, C_t, C_r, and $\mathbf{o}$ are conditionally independent.[11]

That is,

$$O_k \perp C_t, C_r, T | \eta_1, \mathbf{x}, \tag{3}$$

so the compliance and missing data processes are ignorable. One can see how the parameters in $\boldsymbol{\theta}_\mathrm{k}$ are identified in equation (1a). Omitting η_1 in the kth outcome equation would make the combined error term, $\eta_1 + \boldsymbol{\epsilon}_k$, correlated with treatment compliance (T) since by construction η_1 is correlated with T, and this omission would bias all structural parameter estimates.

Since the GET model controls for subjects' compliance type using an estimated latent variable, we assume overlap on this latent variable across the treatment assignment arms $\mathbf{z}$ of the experiment; randomization ensures this is true in expectation. This assumption also requires that the compliance processes measured in $\mathbf{c}$ are not deterministic, so that some subjects with a relatively low probability of complying happen to comply with the treatment. Since by assumption subjects in the control group do not have access to the treatment, randomization ensures that some subjects with a high compliance probability do not receive the treatment.

Because we observe the compliance covariate for everyone in the sample, GET can retrieve the ATE. The ATE for an outcome equation k is,

$$\Delta_k^{\mathrm{ATE}} = \int_{-\infty}^{+\infty} [\Phi(\alpha_{0k} + \alpha_{1k} + \lambda_k \eta_1) - \Phi(\alpha_{0k} + \lambda_k \eta_1)] \phi(\eta_1) d\eta_1. \tag{4}$$

This representation assumes that the covariates have been mean differenced, and one is interested in effect estimates for an average subject in the sample.

The term inside of the square brackets is the estimated difference in potential outcomes conditional on compliance type (below we consider the case of treatment effect heterogeneity, where λ_k can differ across the terms inside of the square brackets). In retrieving a causal estimand, the expectation of the difference is over the unconditional distribution of the compliance variable, $\phi(\eta_1)$, rather than on a conditional distribution such as $\phi(\eta_1 | Z = 1)$. Note that constraining λ_k to zero reduces the GET model to probit, which is biased in the case where $\delta O_k / \delta \eta_1 \neq 0$.

GET also can retrieve other estimands of interest. In particular, we define the complier average causal effect Δ_k^{CACE} as identical to the parameter in equation (4) except the bounds of integration range from 0 to $+\infty$. To be consistent with existing approaches to principal stratification, in the remainder we focus on the complier average causal effect (CACE) estimand.

Assuming latent ignorability and the exclusion restriction, all dependence between the endogenous elements of $\mathbf{e}$ is captured by the latent variable η_1. Including this latent variable in multiple equations allows estimation of the correlations among and between all endogenous compliance variables $\mathbf{c}$ and the outcome variables $\mathbf{o}$, under the assumed DGPs and after conditioning on observed pretreatment covariates. For example, when all outcome and compliance variables are dichotomous, the correlations between each compliance behavior and each outcome can be retrieved with:

[11] Indeed, this form of conditional independence is a standard assumption in the item response theory literature (Trier and Jackman 2008 ; Patz and Junker 1999).

$$\rho o_k, c_m = \frac{\lambda_k \lambda_m}{\sqrt{\left(\lambda_k^2+1\right)\left(\lambda_m^2+1\right)}} \tag{5}$$

for all m, k. Equation (5) follows from the definition of linear correlation and the distributional assumptions for $\{\eta_1, \epsilon_k, \epsilon_m\}$. Equation (5) presents a formal test for dependence that is driven by unobservables that may remain after balancing the sample or conditioning on observables. As we demonstrate below, this enables a test of whether the treatment is ignorable after balancing on observed covariates, such as through matching.

To retrieve the correlations among the compliance variables using equation (5), substitute for $m' \neq m$ for k, and to retrieve the correlations among the outcomes, substitute for $k' \neq k$ for m. Note the former correlations help to assess the validity of the latent measure for compliance. The latter correlations among related dependent variables, such as between math and verbal test scores, are commonly encountered (e.g., Barnard et al. 2003, 305).

3.3 *Extensions*

Here we relax some important assumptions built into the basic GET model of equations (1a) and (2a).[12]

3.2.1 Relaxing the exclusion restriction

In IV analysis, including IV methods based on principal stratification, a violation of the exclusion restriction will bias the CACE estimate. The exclusion restriction if often violated in practice. For example, Hirano, Imbens, and Zhou (2000) analyze data from an experiment testing the efficacy of an influenza vaccine in which physicians were randomly chosen to receive a letter reminding them to vaccinate their patients. They found evidence to suggest that the physicians who received the letter may have taken additional steps to help their patients avoid exposure to the flu, especially if the vaccine were refused. In this case, the effect of the letter could not necessarily be attributed to the efficacy of the vaccine itself.

In contrast, GET is based on a structural equation measurement model (see, e.g., Bollen 1989), in which the compliance type η_1 enters as a latent covariate. If one suspects that any part of the compound exclusion restriction is violated, one can enter the assignment variable Z as an exogenous regressor in the appropriate equation, such as in any of the equations for O_k and C_r.

3.2.2 Treatment effect heterogeneity

GET can test for heterogeneous treatment effects across a range of the compliance type latent variable (e.g., Horiuchi et al. 2007) by estimating each λ_k as an expanded function of the (endogenous) treatment variable, $\lambda_k = \lambda_{1k} + \lambda_{2k}T$. This expansion allows the treatment effect to vary across subjects as a function of their unobserved propensity to take up the treatment (Björkland and Moffitt 1987). In addition, this expansion of λ_k identifies the correlation between the individual's treated and untreated conditions (Aakvik et al. 2005). Setting $\lambda_{2k} = 0$ assumes homogenous treatment effects.

3.2.3 Nonlinearities

The latent variable η_1 is a continuous and unidimensional variable, and it enters equation (1a) linearly. If one suspects that the outcomes are nonlinear in compliance type, one can enter higher order terms. For example, one can enter the second-order term η_1^2 if one suspects there is a diminishing effect across the range of complaince type.

[12] As in any multilevel model, GET is flexible enough to account for complex data structures, including certain violations of the stability assumption. For example, the model can accommodate variations in the distribution of compliance types across experimental sites using a level three random intercept, and the treatment coefficient can vary across sites with a level three random coefficient. The variances of the outcomes also can be estimated when they are identified.

3.4 *Estimation*

In the application below, we estimate the parameters of the $m + k$ equations simultaneously using Bayesian MCMC methods with data augmentation to simulate the posterior distribution (Imbens 1997). Assuming the form of conditional independence implied in latent ignorability, defining $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ to collect all structural parameters, taking $p(\boldsymbol{\theta})$ and $p(\eta_1)$ as priors, and assuming that the pretreatment covariates $\mathbf{o_{pt}}$ are modeled as endogenous variables as in equation (2a), the posterior distribution is

$$p\left(\boldsymbol{\theta}, \eta_1 \middle| \mathbf{o}, T, \mathbf{o_{pt}}, \mathbf{x}, \mathbf{c}\right) \propto p\left(\mathbf{o}|\eta_1, \boldsymbol{\theta}, \mathbf{x}, T, \mathbf{o_{pt}}\right) \times p\left(\mathbf{c}|\eta_1, \boldsymbol{\theta}, \mathbf{x}\right) \times p\left(\mathbf{o_{pt}}|\eta_1, \boldsymbol{\theta}, \mathbf{x}\right) \times p\left(\boldsymbol{\theta}\right) \times p\left(\eta_1\right) \qquad (6)$$

We implement the models below in the MCMC software WinBUGS, which alternately imputes missing values in $\mathbf{o}$ and $\mathbf{c}$ and using estimated parameter values (Tanner and Wong 1987),[13] and then updates the parameter values with the augmented data until convergence (see Jackman 2000). In this approach, one can approximate maximum likelihood (ML) estimates by assigning flat priors for $\boldsymbol{\theta}$, although in some instances assigning informative priors is theoretically justified and/or can aid convergence. In particular, restricting the sign of factor coefficients $\boldsymbol{\lambda}$ may be helpful to aid convergence, and the researcher may have strong a priori beliefs regarding the relationships among the compliance indicators. For example, in the application below, one could argue it is implausible to believe that engaging in an online deliberative session would be negatively related to responding to online surveys.

The GET model differs from standard item response models (e.g., Trier and Jackman 2008) as it includes an indicator of the latent variable, T, as an endogenous regressor in the outcome equations (see equation 1a). Miranda and Rabe-Hesketh (2006) and Terza (1998) demonstrate that no restrictions on the structural coefficients are necessary for theoretical identification of the nonlinear outcome equations. Analytically proving sufficient conditions for identification of complex multilevel models can be intractable. In the ML context, full rank of the estimated information matrix is a necessary and sufficient condition for both theoretical and empirical model identification (Skrondal and Rabe-Hesketh 2004, 150–1). For example, below we describe artificial data used in a Monte Carlo study; in one such data set, the condition number for the ML information matrix at the solution was 17.4, with smallest eigenvalue 2.2. That this estimated information matrix is full rank shows the basic GET model is identified. As in any regression model, GET can be empirically underidentified in a specific application.

3.5 *Monte Carlo Studies*

In the Appendix, we report extensive simulation studies comparing the performance of GET to an implementation of FR (Imai 2009). The simulations demonstrate five advantages of GET. First, GET can estimate treatment effects more efficiently since GET exploits more information to measure compliance type, and this efficiency improves as more compliance indicators are added to the model. Second, GET can test for violations of the exclusion restriction and corrects bias in treatment effect estimates when this assumption is violated. Third, in the presence of treatment effect heterogeneity, GET appears to underestimate treatment effects but still with greater efficiency, for an overall reduction in root mean square error. In addition, GET is able to retrieve direct estimates of treatment effect heterogeneity across the range of the compliance variable. Fourth, GET appears to be as robust as FR to misspecification in the functional relationship between compliance and outcome (in this case, assuming a linear relationship when the true relationship is concave). Fifth, we demonstrate how to implement a discrepancy test described in Barnard et al. (2003).

4 Application: Citizen Efficacy in a Deliberative Experiment with Members of Congress

In this section, we show how to apply GET in practice. In the summer of 2006, we conducted a series of online deliberative field experiments, where current members of the U.S. House of Representatives interacted via a Web-based interface with random samples of their constituents. Twelve members of Congress

[13]The missing data are imputed as missing at random (MAR); MAR follows from the conditional independence assumption of latent ignorability.

conducted either one or two sessions for a total of 22 sessions. The number of participants in each session ranged from 8 to 30 constituents.

The topic of each session was federal immigration and border control policy. During the session, constituents were asked to type questions and comments into a text box, which were then placed in a queue. A moderator screened the postings for redundancy and then sequentially posted the questions and comments to the screen for the member to respond. The member responded to the questions orally, with the audio available on the subjects' computer sound system. Simultaneously a real-time captionist typed the member's responses into a textbox, so constituents could both hear and read the member's responses. After 35 min, the member logged off the session, and the constituents continued the discussion among themselves in an online chat room. These constituent-only discussions typically focused on the member's performance, immigration policy, and the deliberative process itself.

The Congressional Management Foundation[14] recruited the participating Members of Congress. Five Republicans and seven Democrats took part, with good variation across region and gender. One from each party was a party leader, and one from each party voted against their party on recent immigration legislation.

Knowledge Networks (KN), an online survey research firm, recruited the constituents from each congressional district and administered the surveys.[15] After completing a pretest survey, each constituent was randomly assigned to one of three conditions: a deliberative condition that received background reading materials on immigration policy and was asked to complete a survey regarding the background materials (the background materials survey) and to participate in the sessions; an information-only group that only received the background materials and was asked to take the background materials survey; and a true control group. A week after each session, KN administered a follow-up survey to subjects in each of the groups.

For this study, we restrict the sample to the 670 subjects who initially indicated a willingness to participate in the deliberative sessions, and who completed the baseline and background materials surveys. Thus, the treatment effect compares those who read the background materials and participated in the discussions to those who only read the background materials. We make this restriction for two reasons. First, within this subsample, the treatment and control subjects are comparable in that they indicated a willingness to participate in a deliberative session if asked, and they exhibited enough motivation to read the background materials and complete two surveys. Second, the treatment is dichotomous (they either participated in a session or not), which simplifies many of the analyzes below considerably. The GET model is general enough to handle the more elaborate comparisons on the full sample, but in the present context, such complexities would needlessly obscure our main methodological points.

4.1 *Outcomes*

For this analysis, we will examine whether subjects in the deliberative sessions report higher levels of internal efficacy, external efficacy, or both, compared to those who only read the background materials (Acock, Clarke, and Stewart 1985). On the follow-up survey, all subjects were asked ''Please tell us how much you agree or disagree with the following statements'':

1. I don't think public officials care much what people like me think.
2. I have ideas about politics and policy that people in government should listen to.

The first question is a measure of external efficacy, and second question is a measure of internal efficacy. The response categories were ''Strongly Agree,'' ''Somewhat Agree,'' ''Neither Agree nor Disagree,'' ''Somewhat Disagree,'' and ''Strongly Disagree.'' To simplify the analyzes, we dichotomized the responses to create two outcome variables: (1) the *Officials care* variable equals one for subjects who somewhat disagree or strongly disagree with the first question, and zero otherwise and (2) the *Have ideas*

[14]CMF is a nonpartisan, nonprofit, dedicated to assisting Members of Congress to better manage their offices. See http://http://www.cmfweb.org.

[15]KN maintains a probability sample panel of survey respondents that is designed to be representative of the U.S. population. See http://www.knowledgenetworks.com/ganp/index.html for technical details. In many of the districts, the KN panel was not large enough to yield a sufficient number of observations for each treatment arm of the study. In those districts, KN subcontracted with other survey firms to get large enough samples. We control for panel differences in the models reported below.

Table 1 Descriptive statistics

	Full data set		*Matched data set*	
	Mean	*SD*	*Mean*	*SD*
Officials care				
Baseline survey	0.150	0.357	0.131	0.337
Follow-up survey[a]	0.216	0.412	0.215	0.412
Have ideas about politics				
Baseline survey	0.670	0.471	0.683	0.466
Follow-up survey[a]	0.681	0.466	0.663	0.473
Participated in session[b]	0.530	0.500	0.524	0.500
Follow-up survey response	0.787	0.419	0.815	0.388
November survey response	0.645	0.479	0.764	0.425

Note. The model also includes 12 exogenous control variables; see Appendix.
[a]*N*for these two variables is 527, given nonresponse on the follow-up survey.
[b]Among those assigned to treatment (83% of sample randomly assigned to treatment).

about politics variable equals one for subjects who somewhat agree or strongly agree to the second question, and zero otherwise. Some subjects, however, chose not to respond to either question, and for these, the outcomes are missing.

Efficacy is an apt object for our current purposes for both substantive and methodological reasons. Substantively, efficacy is relevant to deliberative democracy for two reasons. On the one hand, deliberative democrats worry about unequal participation in deliberation based on varying levels of efficacy (Young 1990). So we might expect nonrandom treatment compliance and nonresponse based on efficacy. On the other hand, many deliberative democrats argue that deliberation is the cure for low political efficacy, so we might expect significant treatment effects (Morrell 2005). Methodologically, examining changes in both internal and external efficacy is useful to demonstrate the need to account for self-selection into treatments. We have strong a priori expectations that subjects with high levels of efficacy will disproportionately comply with the treatment compared to those with low levels of efficacy, and this should particularly be true for those with high internal efficacy. That is, we have chosen to analyze outcomes where treatment compliance is very likely to be correlated with the potential outcomes. We do not expect, however, to see differences in treatment effects between treatment and control groups for these outcomes since subjects are unlikely to select themselves into the treatment anticipating disproportionate gains in efficacy.[16]

4.2 *Data*

The data are summarized in Table 1 in the column labeled ''Full data set.'' For the outcome variables, the *Officials care* and the *Have ideas about politics* variables, we have both pretreatment responses from the baseline survey ($N = 670$) as well as post-treatment responses from the follow-up survey ($N = 670$ with 143 nonresponses). On both surveys, small percentages of subjects disagreed with the statement that officials do not care what ordinary people think, with only 15% disagreeing on the baseline survey and 22% disagreeing on the follow-up survey. This suggests low levels of external efficacy by this measure. In contrast, a large percentage of subjects indicate relatively high internal efficacy, agreeing that they have ideas about politics and policy that people in government should listen to: 67% agreeing on the baseline and 68% agreeing on the follow-up. Note that there are likely to be many omitted causes that affect each measure of efficacy for each subject on both the pre- and the posttests (see Morgan and Winship 2007, 71). As a consequence, the analyses below treat the baseline outcomes as endogenous variables.[17]

[16]Heterogeneous treatment effects are often observed in experiments where subjects take up the treatment anticipating a larger marginal gain from the treatment. For example, subjects who enroll in a job training program may anticipate a larger wage increase from the program compared to those who choose not to enroll.
[17]Morgan and Winship (2007, 71) demonstrate that matching or conditioning on a pretreatment outcome measure will typically bias treatment effect estimates since this amounts to including a lagged endogenous variable on the right hand side of the equation. Below we show how to include baseline outcome measures as endogenous regressors in GET.

Table 2 Treatment effect and latent correlation estimates

	Full data set				*Matched data det*[a]			
	Officials care (posttreatment)		*Have ideas about politics (posttreatment)*		*Officials care (posttreatment)*		*Have ideas about politics (posttreatment)*	
	Δ^{CACE}	*SE*	Δ^{CACE}	*SE*	Δ^{CACE}	*SE*	Δ^{CACE}	*SE*
Probit[b]	0.087	0.037	0.154	0.047	0.096*	0.028	0.133*	0.034
FR[b]	0.140	0.066	0.030	0.083	—	—	—	—
GET model								
CACE	0.085	0.032	0.005	0.021	0.075*	0.016	0.006	0.019
Latent correlations								
Participated in session	−0.060	0.130	0.420	0.041	0.073	0.039	0.352	0.036
Respond follow-up survey	−0.091	0.198	0.646	0.032	0.141*	0.072	0.688*	0.006
Respond November survey	−0.028	0.063	0.179	0.094	0.090	0.050	0.454	0.120
Officials care (pretreatment)	−0.017	0.041	0.123	0.055	0.011	0.10	0.052*	0.026
Have ideas (pretreatment)	−0.001	0.016	−0.002	0.049	0.006	0.005	0.029	0.017

[a]CACE estimates and latent correlations in the matched data represent weighted averages for matched samples for the treatment and control subjects.
[b]CACE estimates conditioned on all pretreatment covariates listed in Table A1.
*$p < .05$.

We have three measures of the propensity to comply with the study, corresponding to the indicator variables (C_t, C_r, $C_{\neg tr}$). We assigned 83% of this sample to the deliberation arm, of which 53% actually *Participated in a deliberative session*, or $C_t = 1$. For those assigned to the control (information only) condition, C_t is missing. All subjects were administered the follow-up survey, and 79% *responded on the follow-up survey*, or $C_r = 1$. We also administered a survey after the November elections to subjects who completed the follow-up survey, and among those, 65% *completed the November survey*, or $C_{\neg tr} = 1$. We do not use responses to questions on the November survey in these analyses but instead only use whether they completed or did not complete this survey as another indicator of their behavioral propensity to comply with the study. Finally, we created a treatment indicator T that equals one if $ZC_t = 1$ and zero otherwise.

We included a variety of variables in the pretreatment survey that we believed would be the best observed measures of subjects' propensity to participate in the online deliberative session and that might codetermine compliance with the treatment, outcomes, and response on the outcomes (we describe these variables in the Appendix). We use these variables in the GenMatch matching software to create a data set that is balanced on observables (Diamond and Sekhon 2007). See the Appendix for our matching procedures. Applying the GET model to the matched data set (see Ho et al. 2007) allows for a formal test of whether the treatment is ignorable after balancing on observed covariates.

4.3 *Results*

Table 2 reports the results for the CACE estimates. The top portion of the table shows the ATE estimates from the probit, FR and GET models.[18] The bottom portion shows the latent correlations among the choice variables retrieved from the GET model (via equation 3). The left side shows the results from each model using the full data set, and the right side shows the results of the models applied to the matched data set. In addition to the treatment indicator, each model conditions on the full list of pretreatment variables (listed in the Appendix), including the baseline pretreatment outcome.

[18]For the FR estimator, we used Kosuke Imai's R package experiment, version 1.1-0, NoncompLI function, "Bayesian Analysis of Randomized Experiments with Noncompliance and Missing Outcomes Under the Assumption of Latent Ignorability." Documentation is available at http://imai.princeton.edu.

The naive probit model, or the probit model applied to the full data set, retrieves large and precise estimates of the ATE for both outcome variables. That is, by the naive estimate, it appears the treatment increases subjects' levels of both internal and external efficacy. By these estimates, participating in a deliberative session tends to increase the chance that subjects believe officials care about what they think (external efficacy) by 9 percentage points ($p < .05$), and that they themselves have ideas about politics and policy that government officials should listen to (internal efficacy) by 15 percentage points ($p < .05$).

One would be incautious, however, to interpret these estimates as unbiased treatment effect estimates since there is a strong chance that subjects with low initial efficacy are disproportionately refusing the treatment. The correlation retrieved using the GET model and the full data set between participating in the deliberative sessions and the belief that one has ideas about politics and policy that people in government should listen to (internal efficacy) is 0.42 ($p < .001$). In contrast, the latent correlation between complying with the treatment and the measure of external efficacy, officials care about what people like me think, is small, only –0.06 (not significant). These results suggest that subjects are complying with the treatment based on internal but not on external efficacy. This selection process is plausible, suggesting that people choose to participate in the deliberative sessions based on their own internal motivation, rather than on their understanding of how representative institutions and government officials themselves respond to constituent input.

We show in the Appendix that the matched data sets have near-perfect balance on the observed covariates. Having perfect balance on the observed covariates, however, does not ensure identification of the mean treatment effects if there is selection on unobservables. When applied to the matched data, the GET model provides a hypothesis test for the ignorability of the treatment after balancing on observed covariates. In this case, the correlations among the choice variables are nearly identical in the matched data set as they are in the full data set. That is, the observed covariates do a poor job in breaking the dependence between treatment compliance and the measure of internal efficacy. This is despite balancing on variables we believed would be the best predictor of selection into a deliberative experiment, the variables measuring need for cognition (Cacioppo, Petty, and Kao 1984) and need for evaluation (Bizer et al. 2004). Indeed, heretofore, the empirical literature on deliberative democracy has yet to develop a highly predictive model of the propensity to participate in deliberation.

Results from the two models that condition on subjects' latent compliance type, GET and FR, suggest that the deliberative sessions have little effect on subjects' levels of internal efficacy. The GET CACE using the matched data is only –0.005 (or nearly identically zero) and this estimate is not statistically significant. Notice that the naive probit estimate is 77 times larger than the GET estimate using matched data, and further a hypothesis test would reject the null hypothesis with the naive estimates but not with the GET estimates. The FR CACE estimate is also statistically zero, but with a standard error that is four times higher than the GET estimate. This is consistent with the simulation results (see the Appendix) that show that GET is more efficient than FR when covariates do a poor job of predicting compliance.

In contrast, given there is little dependence between complying with the treatment and the measure of external efficacy, the GET, FR, and probit estimates are reasonably similar. By the GET model with the matched data, participating in a deliberative session increases subjects' belief that government officials care about their ideas by about 8 percentage points ($p < .05$). This estimate is statistically equal to the naive probit estimate, and by both models the analyst would be led to reject the null hypothesis.

The GET point estimates with the full data set are nearly identical to those using the matched data. In both cases, the analyst would be led to accept the null hypothesis that the deliberative sessions do not affect internal efficacy and to reject the null hypothesis that the sessions do not affect external efficacy. Overall, the GET model estimates remains robust across the two data sets since GET is accounting for the latent dependencies between treatment compliance, response, and outcome, given that the pretreatment covariates are only weakly related to the compliance process.

Substantively, the differing results found in the GET model are quite sensible. On the one hand, participating in a discussion with others is unlikely to change one's sense of internal efficacy regarding the quality of their own ideas about policy. In these sessions, one could imagine some subjects realizing that their ideas are just as good as others' ideas or perhaps better or worse. On the other hand, having the chance to interact with their member of Congress, and observe the member give unscripted and generally thoughtful responses to participants' questions and comments, appear to have a direct and strong impact on subjects' sense of external efficacy. Participants in the sessions are able to observe directly how a government

official cares about how her constituents think about the immigration issue and how the member treated their questions and the other constituent questions with respect.

The correlations among the indicators retrieved from equation (5) show the validity of the latent compliance measure. The correlation between participating in a discussion and responding on the follow-up survey is 0.54, between participating in the discussion and responding on the November survey is 0.15, and between completing the follow-up survey and responding on the November survey is 0.23, all with $p < .05$.

The GET model also can account for dependence between endogenous regressors (other than the endogenous treatment) and the outcome responses (see Morgan and Winship 2007, 71). We note that pretreatment measures of survey outcomes tend to be codetermined with the post treatment measures (Achen 1975). We can test for such dependency by allowing the pretreatment measures also to load on the latent variable η_1. Here we observe that in this application there is a statistically significant correlation between pretreatment and posttreatment responses on the measure of internal efficacy, but the level of correlation is substantively small. No pre-post correlation is observed for the measure of external efficacy. As a consequence, one could justify including the baseline responses as exogenous control variables.

The GET model can retrieve heterogeneous treatment effects when the heterogeneity is a linear function of the latent propensity to comply with the study. In this application, we discovered no heterogeneity in this functional form using the full data set (results not reported). The GET model also can relax and test the exclusion restriction that assignment has no effect. In these data, one can imagine that assignment might have an effect on participants' attitudes since in this case it is an invitation to meet with their member of Congress. We re-estimated the model including the compliance variable in each outcome equation and in the equation modeling the propensity to respond on the follow-up survey. We find no significant effects on the latter two. We do find evidence, however, that assignment to the treatment enhances subjects' sense of internal efficacy. After accounting for the effect of assignment, the results suggest that actually participating in a session reduces one's sense of internal efficacy an equal amount (hence the net effect of zero CACE estimated above). If this were a substantive study of efficacy, we would examine this interesting pattern more closely.

5 Discussion

In the FR approach, the method of principal stratification can observe only dichotomous realizations of the compliance propensity, C_t, and then only among subjects who are assigned the active treatment. For those assigned to the control group, these realizations are missing data to be imputed through model-based assumptions and measured covariates (e.g., Frangakis and Rubin 2002; Horiuchi et al. 2007). In many applications, however, compliance data on all subjects are collected as a part of an experiment. When such data exist, GET exploits these additional responses to measure the compliance propensity for all subjects. As a result, GET has the potential to improve efficiency as it makes use of all compliance data available; the compliance type is measured rather than imputed for those assigned to the control. Like FR, GET must impute compliance with the treatment for those assigned to the control, but this imputation is conditioned on compliance type, as well as observed covariates. Assuming the compliance indicators are valid and reliable, compliance type by construction is a strong predictor of compliance with the treatment, and hence there is less uncertainty in this missing data process.

We show that exploiting the additional compliance information on all subjects has a number of other benefits. First, GET allows one to relax and test the exclusion restriction that requires the assignment to have no effect on the outcome or the probability of response, a restriction that is often implausible in practice. Second, GET can make treatment effect heterogeneity a function of latent compliance type, allowing the effect of the treatment to vary as a function of the continuous compliance latent variable. And third, GET provides a useful diagnostic to test for the ignorability of treatment compliance after matching.

Implementing GET does require, however, that these additional compliance indicators exist. Often, collecting these data can be built into the experimental design. For example, the study may administer multiple waves of a survey or build in tasks into a multistage experiment. Or, alternatively, the past compliance history of subjects may be recorded, as when using a survey firm that maintains a panel of respondents to whom multiple surveys are administered. Note however that two out of three of our compliance indicators are also required for FR's approach, a partially observed indicator for complying with the

treatment and a fully observed indicator for responding on the outcome. To implement GET, one only need a single additional indicator of compliance.

In our application, we find there is a considerable amount of noncompliance with the deliberative session that is not captured by either the theoretically compelling measures of need for cognition and evaluation or by attribute data. In particular, we find that subjects are selecting into the deliberative sessions in a manner that is correlated with internal efficacy and not with external efficacy. Accounting for these compliance processes matters considerably in assessing treatment effects.

6 Conclusion

In any application where compliance and response are self-selected, there is likely a danger that selection into and out of the study is driven by unobservables. Noncompliance and nonresponse are likely to be a significant problem in any field experiment. GET provides a useful diagnostic tool for applied researchers when there is the suspicion the compliance processes are driven by unobservables and further offers one approach to identify treatment effects in this situation. GET identifies the latent propensity to comply in the study using behavioral measures, rather than attributes or survey responses that often have only modest use in predicting compliance.

Thus, in designing a study, in addition to collecting an extensive set of control variables, researchers who anticipate using an approach like GET should build in behavioral choice opportunities to comply or not comply with the study and to collect any other indicators of compliance that may be available. In this application, we gave subjects multiple opportunities to complete surveys, a behavior that turns out to be strongly correlated with compliance in the experiment. The more of these behavioral measures that the researcher collects, the better she will be able to estimate the latent tendency to comply, which is the core problem in identifying treatment effects in experiments.

Funding

Digital Government Program of the National Science Foundation (award number IIS-0429452).

Appendix

In this Appendix, we first describe a series of Monte Carlo simulations that demonstrate the advantages of GET over FR when the additional compliance indicators exist. We then discuss how we implement matching for our estimation procedures, as well as describe and report balance scores for the covariates we use for matching.

Monte Carlo Studies

We use simulation methods to compare the performance of GET relative to the version of principal stratification described in Frangakis and Rubin (1999)[18] in which the compliance type is unobserved for those assigned to the control. For each simulation, we draw 50 independent samples of 1000 observations each, composed of compliers and never takers. We randomly assign approximately half of the observations to the treatment condition ($Z = 1$). We construct each choice variable $\{O_k, C_t, C_r, C_{\neg tr}\}$ as a vector of binomial draws with linear index composed of exogenous variables ($\mathbf{x}$), an independent, normally distributed ''unobserved'' compliance variable (η_1) with mean zero and variance one, a treatment indicator (T), and a fixed coefficient vector. The exogenous variables are three multivariate normal covariates with mean vector zero and a sigma matrix with ones on the diagonal and 0.3 on all off diagonals. We construct the treatment indicator as $T = C_t \times Z$ if $Z = 1$ to account for noncompliance among those assigned to the treatment condition and $T = 0$ if $Z = 0$ to reflect our strong monotonicty assumption. We set compliance with the treatment (C_t) to missing when $Z = 0$, and the outcome to missing when $C_r = 0$. In each case, we set the probability of response to approximately 0.80.

Each model is then applied to each simulated data set to estimate posterior parameter distributions, using the means of the posterior distributions as point estimates. The model specifications exclude the

''unobserved'' compliance latent variable and include the treatment indicator. Omitting the compliance latent variable in a naive model will induce a correlation between the outcomes, the choice to comply with the treatment (when offered), and the missing data pattern; hence the naive model that conditions only on observables will overstate the effect of the treatment since T by construction is strongly correlated with the unobserved variable η_1 by way of C_t. For each choice variable, each coefficient vector gives about 20 times more influence to the unobserved compliance variable than to the exogenous variables, in order to illustrate the strengths of GET when it is most appropriate: when the observables are weakly related to the outcomes and to the compliance process. It is in this situation that having additional information on compliance type is most beneficial.

Efficiency when the exclusion restriction is met

We first analyze the relative efficiency of GET and FR when the exclusion restriction is met. Both approaches impute compliance with the treatment for those assigned to the controls and propagate the uncertainty of this imputation through all estimated parameters (Tanner and Wong 1987), but each model treats this (partially) imputed variable in very different ways. FR takes C_t as itself a measure of compliance type, and this variable enters the outcome equation as a covariate. In FR, the imputation of compliance is based only on the exogenous variables, which in the simulation are only weakly predictive of choices, so the imputation of compliance type among those in the control should have a considerable amount of uncertainty.

In contrast, GET uses the estimate of the latent variable η_1 as the measure of compliance type, and C_t is only one of several indicators, which in this case are all highly correlated. Each observation has some information on the compliance type since two of the three compliance indicators are observed for each subject. As a result, GET should estimate compliance type with a considerable degree of certainty. In addition, although GET also must impute C_t for those assigned to the control, it does so based on both the covariates and the latent measure of compliance type. Since the latter is a powerful predictor of C_t, this imputation should occur with relatively low uncertainty.

Although having three indicators is a sufficient condition to identify a latent variable (Bollen 1989, 244), it is possible that adding additional indicators will improve measurement of the latent variable. To test this, we implement GET in two ways, the ''low information'' version that uses only three indicators of compliance, $\{C_t, C_r, C_{\neg tr}\}$, where the last term is a scalar and a ''high information'' version that instead has five elements in $\mathbf{C}_{\neg tr}$.

To draw the outcome variable, we set the coefficient on T to zero (so the true CACE estimate is scaled to zero), the coefficient on the unobserved compliance variable to 2, and the coefficients on the exogenous variables to 0.1. The point estimate distributions for all three models are diagrammed in Fig. A1, which shows the point estimates of the CACE rescaled into percent changes. Since the treatment effect is zero, this figure also graphs the bias in the estimates across samples. Both GET and FR are unbiased across samples, although much of the mass of the FR estimates are to the right of zero. Notice there is considerably more density in the tails of the histogram for FR compared to the two GET models. The standard

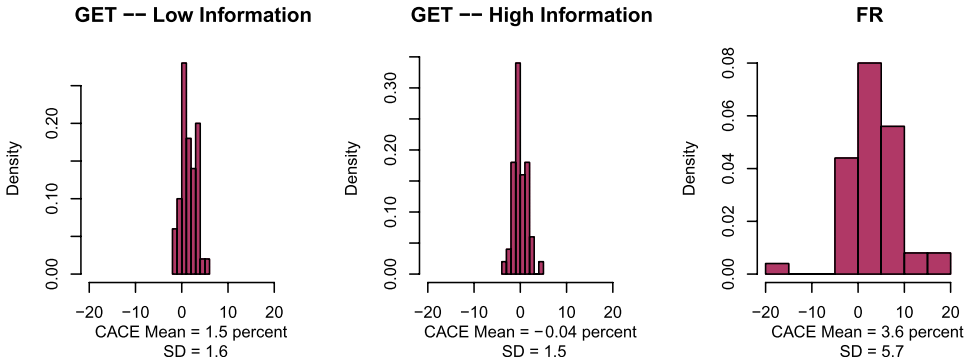

Fig. A1 CACE estimate distributions.

deviations for FR are about 3.6 times higher than for either GET model. The point estimates for the two GET models appear to have similar variability, although there is a slight (but not significant) reduction in bias with the high information model.

This difference in the efficiency of FR and GET is shown in Fig. A2. This figure uses quantile-quantile (QQ) plots to compare the distributions of the root mean squared error (RMSE) of the point estimates of the different estimators. In the graphs marked "A" and "B," one can see the thicker tails of the FR CACE point estimate distribution compared to GET as the points all fall above the 45 degree line. In the figure marked "C," one can see that GET's RMSE is lower in the high information than in the low information case.

Correcting bias when the exclusion restriction is violated

We next explore the relative performance of GET and FR when the exclusion restriction is violated. This simulation is identical to the previous one with the exception that we include the assignment variable (Z) in the linear indexes for both the outcome and the response equations with coefficient equal to one (or 10 times the impact of each exogenous covariate), and we reduced the impact of the unobserved compliance variable to only five times that of each exogenous covariate. We used the low information version of GET. As before, we set the treatment effect to zero. The results of this simulation are presented in the first two panels of Fig. A3, where again these plots show the degree of bias from the true effect.

The difference in the performance of the two models is dramatic, but not unexpected. In effect, the FR model omits the main causal variable, assignment, and hence the results exhibit considerable omitted variable bias. FR's standard deviation remains about four times higher than that of GET. Given the bias and the inefficiency of FR in this case, the QQ plot of the two RMSE distributions, shown in the third panel of Fig. 3, is even more dramatic in the comparison.

Treatment effect heterogeneity

We demonstrate how GET can estimate treatment effect heterogeneity in this simulation by drawing the outcome variable as a function of the unobserved compliance variable, the treatment indicator, and the interaction of the two, each at five times the impact of the covariates. In this case, we used the high information model as it converges more readily than the low information model (which suggests an increase in the performance of the model with additional indicators, in the case where the latent variable is more than a simple random intercept). Figure A4 shows the bias in the overall CACE for each model, where we used the fixed coefficients for the linear index to compute the benchmark CACE and then we subtracted this true value from the estimated value for each iteration. The first two panels of this figure shows that GET continues to exhibit a greater efficiency, but in this case FR appears to be less biased. The third panel of the figure shows that GET's RMSE is equal to or better than FR's across all data sets, indicating that the trade-off between bias and efficiency favors GET in this simulation, although not dramatically.

The strength of GET in this example is in its ability to estimate treatment effect heterogeneity across a range of the compliance variable. The CACE in GET is calculated by integrating over the support of

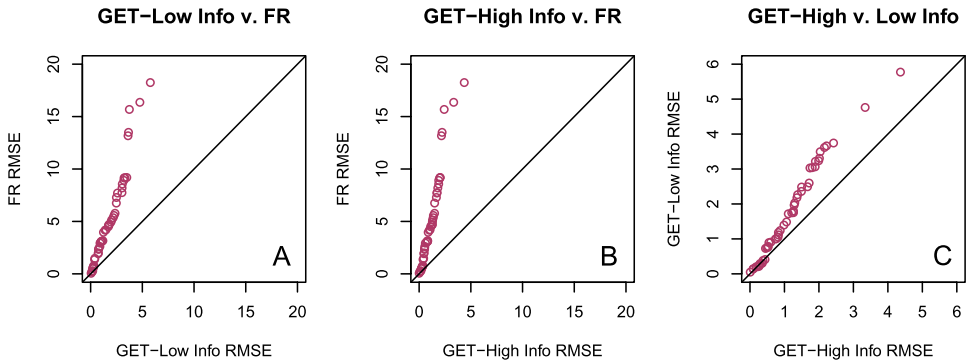

Fig. A2 CACE MSE QQ plots.

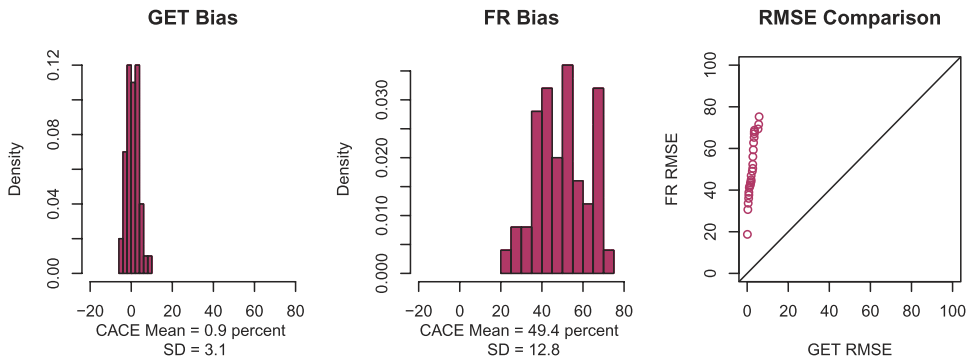

Fig. A3 CACE estimates when exclusion restriction is violated.

likely compliers on the latent variable η_1. GET can retrieve conditional distributions of the treatment effect at any point on the support of the compliance variable simply by holding the compliance variable fixed. In this example, we find that the treatment effect is approximately 1.5 times larger for those who score one standard deviation above the mean on the compliance variable compared to those who score at the mean (although the difference in these point estimates is not statistically significant).

Nonlinearities

The previous simulations assume the researcher already knows the form of the relationship between compliance type and the outcome. And not surprisingly, the model performs quite well in these cases where the model exactly matches the structural features of the DGP. As in any econometric model, GET assumes the researcher has strong theoretical knowledge of the model specification, along with functional and distributional forms. GET possesses the flexibility inherent in any Bayesian model and can accommodate nonparametric priors, multilevel dependence, heteroskedastic variances (when identified), and so on, to match the modeling assumptions to the rich theoretical understanding a researcher may have regarding the DGP.

That said, it remains worthwhile to explore the implications for inference when the model is incorrectly specified. Although there are many ways to mis-specify a model, we explore the consequences when the modeler assumes the outcome index is linear in compliance when it is in fact concave (diminishing). To evaluate this, we sampled the outcome variable using a linear index with the impact of the unobserved compliance variable positive and about three times the impact of the covariates and the impact of a second-order term of the compliance variable about the same as the covariates. This relationship assumes a positive but diminishing relationship between compliance and the outcome linear index, with a negative relationship at very high values of compliance.[19] We again set the treatment effect to zero.

We estimate a ''first-order'' GET that enters only η_1 (restricting the relationship to be linear) and FR. Because the model has a difficult time converging, we only estimate these models on 10 independent data sets using the ''high information'' version of GET. In this case, each of the three models is unbiased and with similar precision. The first-order GET model has a bias of about 2.8% (SD of 6.5%), and FR has a bias of about 1.3% (SD of 6.9%). We also estimated a ''second-order'' GET model that enters both η_1 and in the outcome index, allowing a concave relationship, but the model only converged three times in our trials. This model had the lowest bias (less than 1%) and was most efficient (SD less than one), but three trials are not enough to fully assess this model.

To show that the unbiasedness for these models is not simply an artifact of the nonlinear relationship between compliance and the outcome, we estimated a naive probit model that omits compliance, which had bias of 13.3% (SD 6.6%, significantly different from the true treatment effect of zero).

[19]We chose to have a mode in the support of the compliance variable as this presents a harder case for the linear assumption than when the relationship is monotonic throughout.

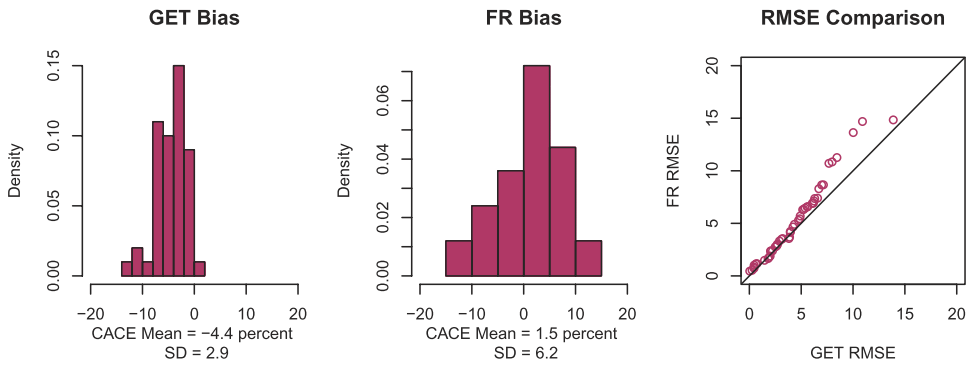

Fig. A4 CACE with treatment effect heterogeneity.

Assessing model fit

Barnard et al. (2003) recommend using a discrepancy test to assess the quality of the model based on the observed and replicated data. To conduct the test, one first uses the simulated posterior distribution of the parameters to replicate the data a large number of times. One can compare some discrepancy measure based on the observed data to the distribution of this statistic across the replicated data. As a discrepancy statistic, Barnard et al. (2003) use the overall intention to treat estimand $\mathrm{E}[O_k(Z=1) - O_k(Z=0)]$ divided by its standard error, with casewise deletion of missing data. For this example, we use the results from the first- and second-order models for the model in the *Nonlinearities* section above. We drew 999 replicated data sets for each model, one for each set of posterior parameters, by setting η_1 to its expected value for each subject, calculating each linear index, drawing the outcome O from a binomial distribution, drawing an indicator of whether the subject responded from a binomial distribution, C_R, and then set the outcome to missing for those with $C_r = 0$. We find the overall discrepancy statistic for the observed data is 0.46. The mean discrepancy statistics for the first-order model is 0.38 and for the second-order model is 0.37. About 46% of the simulated distribution lies above 0.46, close to 50% that one would expect, although it is possible that this implementation of GET results in a somewhat conservative estimator.

Combining GET with Matching

Nonparametric matching methods have been proposed to relax model-based parametric assumptions (see Abadie et al. 2001; Ho et al. 2007; Imbens 2004; Rosenbaum and Rubin 1985). Among its virtues, matching does not assume a functional form for the treatment effect itself nor does it assume a functional form for any heterogeneity in the treatment effect. In addition, matching explicitly calls attention to the problem of comparability between treatment and control subjects and helps to identify which cases are outside the common support of the data.

In many cases, however, matching's assumption of selection on observed variables may be much stronger than the assumptions embedded in parametric models (see Morgan and Winship 2007, 77). In practice, it is quite common to have compliance based on unobserved measures, particularly in any study when subjects self-select their treatment compliance or their response behavior on the outcome measurement. In the case of unobservables, matching simply substitutes one set of strong assumptions for another. Parametric models require some faith in the relative robustness of the distributional and functional form assumptions, whereas matching models require the hope that selection depends only on observed covariates. On their own, either set of assumptions is likely to cause skepticism among the intended consumers of the analysis.

As a way out of this predicament, Ho et al. (2007) propose using matching to preprocess data for use in a parametric model. As they note, parametric methods allow for sample selection on exogenous variables without biasing the estimates of structural parameters. Thus, an analyst can use matching procedures to create a data set where the observations are weighted such that the observed covariates are balanced. Applying parametric methods to preprocessed data yield results that are "doubly robust" (Ho et al. 2007, 215).

If the assumption of selection on observables is met, and there are comparable cases along the propensity score for each group, then the treatment will be ignorable with matching alone. In this case, the treatment effect estimate will not depend on model specifications or functional forms, and any parametric method will retrieve the ATE (Ho et al. 2007, 212). Since the cross equation correlations estimated in GET test for the existence of dependence, applying GET to preprocessed data provides a formal hypothesis test for whether receiving the treatment is ignorable after balancing on observed covariates. That is, GET can use equation (5) to test whether endogenous selection remains an issue whether or not balance has been achieved among the observed covariates. If the treatment is not ignorable by this test, then GET may still identify treatment effects from the matched sample within the assumptions of latent ignorability.

Combining GET with matching relaxes the assumption that treatment effect heterogeneity is a linear function of the compliance covariate. The average treatment effect for the treated (ATT) is identified when the ATE function is applied to the structural parameters retrieved from a balanced data set where control subjects are matched to treatment subjects (the treatment matched data), and the average treatment effects for the controls (ATC) is identified when the ATE function is applied to the structural parameters retrieved from a balanced data set where treatment subjects are matched to control subjects (the control matched data). The overall ATE is π(ATT) + (1 $-$ π) (ATC), where π is the rate of compliance with the treatment.

Given the large number of exogenous stratifying variables, we construct a propensity score model to use for matching. In the model, we included all exogenous variables listed above (the need for cognition and need for evaluation variables, and the attribute variables) as well as a fixed effect for each congressional district in a logit model and retrieved the propensity score using the estimated linear index function. For the matching algorithm, we called the GenMatch software (Diamond and Sekhon 2007) from within the R program MatchIt (Ho et al. 2004) with the population parameter set to 1000. GenMatch searches the distance metric space to find the metric that minimizes the maximum imbalance in the set of observed covariates (Diamond and Sekhon 2007, 9). We matched on the propensity score and the attribute variables.

We used MatchIt and GenMatch to create two preprocessed data sets. For the first data set, we used the naturally coded treatment indicator (1 if treatment, 0 if control) to optimally match controls to treatments. In this data set, 341 of the original 373 controls were matched (with 2 treated and 1 control discarded). For the second data set, we used the reverse coding of the treatment indicator (1 if control, 0 if treatment) to optimally match treatment subjects to controls. In this data set, 181 of the original 297 treatment subjects were matched. A comparison between treatment and control in the first data set, the ''treatment matched'' data set, identifies the ATT, and those between control and treatment in the second data set, the ''control matched'' data set, identifies the ATC. A weighted average of these two estimands yields the ATE, with the weight given by the overall treatment compliance rate (0.533).

The GenMatch matching algorithm improves balance in the exogenous covariates in two ways. First, GenMatch discards both treatment and control subjects that are not in the common support of the propensity score. In this data set, there were only two observations outside of the common support. Second, GenMatch creates a set of weights which, when applied in a subsequent analysis, optimally matches the distributions of the covariates among the remaining observations, giving nonzero weight only to observations that are retained as matches. In the treatment matched data set, all treatment observations are given the weight of 1, and weights for the matched controls had a mean of 2.82 and a standard deviation of 3.73. In the control matched data set, all control subjects have a weight of 1, and the matched treatment subjects have weights with a mean of 1.95 and a standard deviation of 1.54.

As we describe in the text, we use an estimation approach (a Bayesian model implemented in WinBUGS) that does not allow weights. When using estimators that do not take weights, one must create simulated balanced data sets, where the simulation selects observations based on their weights assigned in a matching algorithm. We first created two weighted-resampled matched data sets, one for the treatment matches and one for the control matches. To simulate the weighted matched data set for the treatment subjects, we first retained all treatment subjects. We then sorted the matched controls by their weights, and assigned each control subject an interval of a line segment with length 1, with the interval length assigned to each control subject equal to the proportion of her weight divided by the sum of all weights. In this interval, unmatched controls have zero length, controls with small weight have a small length, and those with large weight have a large length.

We then drew 297 random numbers (equal to the number of original control subjects) from a uniform [0,1] distribution with replacement, and for each draw, we sampled the control subject whose line segment contained the number of that draw. We used the analogous steps to construct the weighted control matched data set. In both cases, we created a series of 10 simulated data sets. Then among these data sets, we retained the data set that had the best balance, given the natural sampling variation inherent in the simulation. We retain the best balanced data set, rather than a random data set, since standard practice is to retry matching until a balanced data set is produced (Ho et al. 2007).[20] To create the correctly weighted matched data set we report in Table 1, for each pair of observations in the treatment and control matched data sets, we drew a binomial variable with $p = 0.533$ and selected the observation from the treatment matched data set if the variable equaled one and from to control matched data set if the variable equaled zero.

Exogenous Variables Included in the Model

We have a total of 12 pretreatment (exogenous) variables, and these are summarized in Table A1. We included two variables on the pretreatment survey that measure subjects' need for cognition (Cacioppo, Petty, and Kao 1984). We asked subjects, ''Would you say you have opinions about . . .'' with response categories ''Almost everything,'' ''About many things,'' ''About some things,'' or ''About very few things.'' To create the Have opinions variable, we coded those who report having opinions about almost everything or many things as one, otherwise zero, with 77% reporting having opinions. We also asked subjects, ''Some people like to have responsibility for handling situations that require a lot of thinking, and other people don't like to have responsibility for situations like that. Do you . . .'' with response categories ''Like them a lot,'' ''Like them somewhat,'' ''Neither like nor dislike,'' ''Dislike them somewhat,'' or ''Dislike them a lot.'' To create the Likes responsibility for thinking variable, we coded those who like these situations a lot or somewhat as one, zero otherwise, with 69% liking this sort of responsibility.

Table A1 Exogenous variable descriptive statistics and balance scores

	Full data set			*Matched data set*		
	Mean	*SD*	*Balance*	*Mean*	*SD*	*Balance*
Need for cognition						
Have opinions	0.773	0.419	0.202	0.776	0.417	−0.025
Likes resp. for thinking	0.685	0.464	0.192	0.667	0.472	−0.035
Need for judgment						
Important to hold opinions	0.781	0.414	0.064	0.761	0.427	0.063
Not neutral about issues	0.654	0.476	0.190	0.656	0.475	−0.089
Not employed	0.463	0.499	0.022	0.490	0.500	−0.056
Completed some college	0.391	0.488	−0.080	0.387	0.488	0.001
Completed college or more	0.451	0.498	0.165	0.444	0.497	0.031
Female	0.661	0.474	−0.103	0.722	0.448	−0.016
White	0.837	0.369	0.027	0.847	0.360	0.026
March session	0.103	0.304	0.145	0.107	0.309	0.256
KN panelist	0.496	0.500	0.162	0.512	0.500	0.122
High political knowledge	0.676	0.468	0.159	0.683	0.466	−0.079

[20] An alternative would be to re-estimate each of the models below with all 10 data sets and then averaging the distributions of results.

Two other variables measure subjects' need for evaluation (Bizer et al. 2004). On the pretreatment survey, we asked subjects, ''Please tell us how much the statement below describes you'': and presented two statements. The first statement was ''It is very important to me to hold strong opinions.'' To create the Important to hold opinions variable, we coded those who answered either ''Extremely'' or ''Somewhat'' characteristic as one, and those responding ''Extremely'' or ''Somewhat'' uncharacteristic or ''Uncertain'' as zero, with 78% reporting that holding opinions is important. The second statement was, ''I often prefer to remain neutral about complex issues.'' To create the Not neutral about complex issues variable, we coded those who reported this as either ''Extremely'' or ''Somewhat'' uncharacteristic as one, zero otherwise, with 65% reporting a preference not to be neutral about complex issues.

We have a number of other exogenous variables measuring various attributes of subjects that may be related to the compliance process: 46% of subjects were Not employed, 39% had Completed some college, 45% had Completed college or more, 66% were female, 84% White, 10% were in a March (expert) session, 50% came from a representative KN panel as opposed to an opt-in panel that KN subcontracted with, and 68% were able to answer at least four of the ''Delli Carpini and Keeter five'' items correctly on the baseline survey (Deli Caprini and Keeter 1993) indicating high political knowledge.

Table A1 reports the descriptive statistics for the matched data set, as well as the univariate balance scores for each exogenous variable for each data set. The balance score is calculated as the difference in the average between treatment and control groups in the relevant data set, divided by the square root of the sum of the within group variance in the full data set (Rosenbaum and Rubin 1985). Overall, the balance is much better in the matched than in the full data set, reflecting the strong performance of GenMatch. Out of 12 covariates, the balance in absolute value improved on eight and worsened on two, with the average imbalance reduced by about 0.06. The original data set is reasonably balanced, perhaps reflecting some success in our efforts at randomization.

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil. 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* 125:15–51.

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens. 2001. Implementing matching estimators for average treatment effects in stata. *The Stata Journal* 1:1–18.

Achen, C. H. 1975. Mass political attitudes and the survey response. *American Political Science Review* 69:1218–31.

Acock, A. C., H. D. Clarke, and M. C. Stewart. 1985. A new model for old measures: A covariance structure analysis of political efficacy. *Journal of Politics* 47:1062–84.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91:444–55.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin. 2003. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York city. *Journal of the American Statistical Association* 98:299–323.

Bizer, G. Y., J. A. Krosnick, A. L. Holbrook, S. C. Wheeler, D. D. Rucker, and R. E. Petty. 2004. The impact of personality on cognitive, behavioral, and affective political processes: The effects of need to evaluate. *Journal of Personality* 72:995–1027.

Björkland, A., and R. Moffitt. 1987. The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics* 69:42–9.

Bollen, K. A. 1989. *Structural equations with latent variables*. New York, NY: John Wiley & Sons, Ltd.

Cacioppo, J. T., R. E. Petty, and C. F. Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48:306–7.

Delli Carpini, M. X. and S. Keeter. 1993. Measuring political knowledge: Putting first things first. *American Journal of Political Science* 37:1179–206.

Diamond, A., and J. S. Sekhon. 2007. *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies* Department of Political Science typescript, University of California, Berkeley.

Esterling, K. M., M. A. Neblo, and D. M. J. Lazer. 2011. Replication data for: ''Estimating Treatment Effects in the Presence of Noncompliance and Nonresponse: The Generalized Endogenous Treatment Model''. http://hdl.handle.net/1902.1/15619 UNF:5:6fj98k0/hbVkkWd2ouG35w== Murray Research Archive [Distributor] V1 [Version].

Frangakis, C. E., and D. B. Rubin. 1999. Addressing complications of intention-totreat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86:365–79.

Frangakis, C. E., and D. B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58:21–9.

Hirano, K., G. W. Imbens, and D. B. R. X.-H. Zhou. 2000. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1:69–88.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2004. *Matchit: Matching as nonparametric preprocessing for causal inference.* Technical report. Cambridge, MA: Harvard University. http://gking.harvard.edu/matchit/ (accessed March 28, 2011).

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

Holland, P. W. 1986. Statistics and causal analysis. *Journal of the American Statistical Association* 81:945–60.

Horiuchi, Y., K. Imai, and N. Taniguchi. 2007. Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science* 51:669–87.

Imai, K. 2009. Experiment, version 1.1-0, noncompli function, Bayesian analysis of randomized experiments with noncompliance and missing outcomes under the assumption of latent ignorability. Documentation available at http://imai.princeton.edu (accessed March 28, 2011).

Imai, K., G. King, and E. A. Stuart. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A: Statistics in Society* 171:481–502.

Imai, K., and T. Yamamoto. 2010. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science* 54:543–60.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 86:4–29.

Imbens, G. W., and D. B. Rubin. 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* 25:305–27.

Jackman, S. 2000. Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science* 44:369–98.

Mealli, F., G. W. Imbens, S. Ferro, and A. Biggeri. 2004. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* 5:207–22.

Mealli, F., and D. B. Rubin. 2003. Commentary: Assumptions allowing the estimation of direct causal effects. *Journal of Econometrics* 112:79–87.

Miranda, A., and S. Rabe-Hesketh. 2006. Maximum likelihood estimation of endogenous switching and sample selection model for binary, count, and ordinal variables. *The Stata Journal* 6:285–308.

Morgan, S. L., and C. Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

Morrell, M. 2005. Deliberation, democratic decision-making and internal political efficacy. *Political Behavior* 27:49–69.

Patz, R. J., and B. W. Junker. 1999. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* 24:342–66.

Rosenbaum, P. R., and D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39:33–8.

Rubin, D. B. 1974. Estimating casual effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall.

Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82:528–40.

Terza, J. V. 1998. Estimating count data with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* 84:129–54.

Trier, S., and S. Jackman. 2008. Democracy as a latent variable. *American Journal of Political Science* 52:201–17.

Young, I. M. 1990. *Justice and the politics of difference*. Princeton, NJ: Princeton University Press.