

Estimating Treatment Effects in the Presence of Noncompliance and Nonresponse: The Generalized Endogenous Treatment Model¹

Kevin M. Esterling
UC–Riverside
kevin.esterling@ucr.edu

Michael A. Neblo
Ohio State University
neblo.1@osu.edu

David M.J. Lazer
Harvard University
David.Lazer@harvard.edu

August 14, 2008

¹This project is funded by a grant from the Digital Government Program of the NSF (award number IIS-0429452). 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 (NSF). We thank Alberto Abadie, Janet Box-Steffensmeier, Bear Braumoeller, Kosuke Imai, Luke Keele, Gary King, William Minozzi, Jas Sekhon, and Craig Volden for very helpful comments. An earlier version of this paper was presented at the Annual Meetings of the American Political Science Association, Chicago, Ill., Aug. 30 to Sept. 2, 2007, and at the Program on Network Governance, Kennedy School of Government, Harvard University, May 2007. For all program code, documentation and sample data, go to <http://ps-experiments.ucr.edu/resources/>

ABSTRACT

If ignored, non-compliance with a treatment and nonresponse on outcome measures can bias estimates of treatment effects in a randomized experiment. To identify treatment effects in the case where compliance and response are conditioned on unobservables, we propose the parametric generalized endogenous treatment (GET) model. As a multilevel random effect model, GET improves on current approaches to principal stratification by incorporating behavioral responses within an experiment to measure each subjects' latent compliance type. We use Monte Carlo methods to show GET has a lower MSE for treatment effect estimates than existing approaches to principal stratification that impute, rather than measure, compliance type for subjects assigned to the control. In an application, we use data from a recent field experiment to assess whether exposure to a deliberative session with their member of Congress changes constituents' levels of internal and external efficacy. Since it conditions on subjects' latent compliance type, GET is able to test whether exposure to the treatment is ignorable after balancing on covariates via matching methods. We show that internally efficacious subjects disproportionately select into the deliberative sessions, and that matching apparently does not break the latent dependence between treatment compliance and outcome. The results suggest that exposure to the deliberative sessions improves external, but not internal, efficacy.

1. INTRODUCTION

When conducting randomized experiments, researchers typically want to identify and estimate the causal effect of a treatment on some measured outcome. For each subject in the experiment, this treatment effect is defined as a counterfactual, or the difference between her outcome in the condition where she received the treatment and her outcome in the condition where she did not receive the treatment (Rubin, 1974). Since it is not possible to observe a subject in both states, assumptions are required to identify this counterfactual for estimation.

If one can assume that those who received the treatment are in every other way comparable to those who did not receive the treatment, exposure to the treatment can be taken as exogenous in the analysis, and one can estimate the average treatment effect (ATE) across subjects through a simple comparison of treatment and control group averages. To take exposure to the treatment as exogenous requires assuming that subjects in both groups have similar potential outcomes in the control state, would have the same response to the treatment if they were exposed to it, and have the same probability of response on the outcome measurements of interest (see Morgan and Winship, 2007, 46). With treatment noncompliance and nonresponse on outcome measures, however, treatment and control group subjects are not identical in their potential outcomes or treatment responses, even with randomized assignment and large numbers of subjects. In the presence of non-compliance and non-response, identifying treatment effect estimates requires additional assumptions.

Nonparametric matching can identify treatment effects assuming that the compliance processes can be captured fully using observed variables (Abadie, Drukker, Herr, and Imbens, 2001; Imbens, 2004; Rosenbaum and Rubin, 1985). In many situations, however, the compliance process is driven by unobserved variables, and in such cases matching cannot identify treatment effects. Often it is difficult to describe or measure subjects'

motivation to comply 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.

To handle the case of compliance processes with unobservables, we propose the generalized endogenous treatment model (or the GET model) to identify treatment effect estimates within a set of parametric assumptions. Building on the random effect approach in Aakvik, Heckman, and Vytlacil (2005), the GET model extends principal stratification (Barnard, Frangakis, Hill, and Rubin, 2003; Frangakis and Rubin, 1999, 2002; Horiuchi, Imai, and Taniguchi, 2007; Mealli, Imbens, Ferro, and Biggeri, 2004) by estimating each subject’s latent compliance type in a measurement model. GET measures compliance type both for those assigned to the active treatment and those assigned to the control. Under the assumption that treatment compliance, response on outcome measures and the outcomes are independent within strata of a measured compliance variable, an assumption known as latent ignorability (Frangakis and Rubin, 1999), conditioning on the measured compliance variable identifies treatment effects in this context.

When multiple indicators of compliance exist, GET can improve on the efficiency of current applications of principal stratification since it imputes compliance with the treatment among the controls based on the latent compliance type as well as covariates, not simply on measured covariates. By construction, the latent compliance type is a powerful predictor of compliance with the treatment, and conditioning on this latent characteristic helps to reduce uncertainty in the imputation of compliance among those assigned to the control group. In addition, GET retrieves quantities of interest not retrieved in standard approaches to principal stratification. First, because it is based on an item response model, GET estimates the full distribution of compliance type for each subject. Second, GET retrieves the correlations between treatment compliance, outcome response, and

the outcomes of interest, which allows tests of treatment ignorability after balancing on measured covariates. Third, the GET model can estimate heterogeneity in the treatment effect across compliance types and it can incorporate endogenous regressors.

In this paper, we implement the GET model using Bayesian estimation methods (e.g., Horiuchi et al., 2007; Imbens, 1997) and demonstrate its properties. First, we use a Monte Carlo study to show GET retrieves benchmark treatment effect estimates in the presence of unobservables, and does so with a lower mean squared error (MSE) than existing principal stratification methods that assume compliance type is observable among, and only among, those assigned to the treatment (Frangakis and Rubin, 2002). Second, using data from a recent deliberative field experiment, we compare the GET results to those produced by non-parametric matching methods for estimating changes in political efficacy. In the application, we formally show that dependence between compliance and outcomes remain in the sample even after balancing on covariates. Once the latent dependence between internal efficacy and treatment compliance is accounted for, exposure to the deliberative session appears to improve subjects' external, but not internal efficacy. Matching alone cannot overcome the bias of the naive estimator, while GET estimates similar treatment effects given either matched data (Ho, Imai, King, and Stuart, 2007) or the full data set. This comparison demonstrates how GET can be used as a diagnostic tool to test formal hypotheses about the extent to which dependence might remain a problem even after balancing on observed covariates with matching.

2. THE GET MODEL

For an experimental subject, the treatment effect is the difference between the outcome responses she would have if exposed to the treatment and the outcome she would have if not exposed to the treatment (Rubin, 1974). That is, we wish to compare potential outcomes from the experiment, $\Delta = E(O_{1k} - O_{0k})$, where O_{1k} is defined as the k th

outcome that would be observed if the subject received the treatment, and O_{0k} is the outcome that would be observed if the subject did not receive the treatment. Note that treatment effects are defined as a function of a subject's potential outcomes in both the treatment state and the control state. Since the same person cannot simultaneously exist in both states, the treatment effect counterfactual is defined in part as a function of missing data, and so it is not identified for estimation without further assumptions.

Typically, researchers attempt to identify this counterfactual by creating treatment and control groups through random assignment. In many experimental settings, however, researchers cannot compel subjects assigned to the treatment condition to comply with the treatment, or subjects in either group to respond on the outcome measurement. Often those who do comply on treatment and response are different in their potential outcomes from those who do not comply. In this case, treatment and control groups are not directly comparable and some statistical adjustment will be required to identify causal treatment effects (Morgan and Winship, 2007, 46, 78).¹

The method of instrumental variables (IV) addresses the problem of noncompliance when the compliance process itself is unobservable (Angrist, Imbens, and Rubin, 1996). IV remains inconsistent in the presence of nonresponse, however, as compliers tend to have a different probability of response than noncompliers (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 Barnard et al., 2003; Frangakis and Rubin, 2002; Horiuchi et al., 2007; Mealli et al., 2004). The Frangakis and Rubin (FR) approach to principal stratification identifies treatment effects by assuming that the outcome is independent of both treatment compliance and the response pattern within strata of a categorical compliance variable, an assumption they label “latent ignorability.” These

¹Two further assumptions often are evoked as well. The stable unit treatment value assumption (SUTVA) simplifies analyzes by assuming the potential outcomes are not themselves a function of the potential outcomes of others exposed to the treatment (Morgan and Winship, 2007, 37). And, the assumption of random sample selection (Imai, 2007, 4); if this latter condition is not met, then only the sample average treatment effect is identified.

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.²

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 the assumptions of latent ignorability, treatment exposure is identified as a causal effect within strata of the compliance variable (Barnard et al., 2003; Frangakis and Rubin, 1999). By conditioning on compliance type and evoking the assumption of latent ignorability, principal stratification makes the treatment assignment, treatment exposure, and the response pattern conditionally independent of the outcome, and hence ignorable. As a consequence, treatment effects are identified even when latent dependencies exist among compliance, response and the outcomes of interest, that is, even in the case of compliance processes driven by unobservables. In particular, the effect of the treatment among the compliers is a local average treatment effect known as the complier average causal effect (CACE).

In the FR approach, however, principal stratification retains the significant limitation that it must treat compliance type as missing data for those assigned to the control group (or the passive treatment), since those assigned to the control do not have the opportunity to reveal their compliance type. In this approach, the compliance type often is imputed using data augmentation in a Bayesian parametric model conditioned on measured covariates (Barnard et al., 2003; Horiuchi et al., 2007).³ Data augmentation procedures that

²One other compliance category logically exists, the “defiers” who take up the treatment if and only if they are not assigned to the treatment, but studies typically assume this category is never observed under a monotonicity assumption.

³Another approach to principal stratification is to evoke assumptions regarding the response behavior of noncompliers under unobserved realizations of the treatment assignment in order to identify elements of a likelihood function (Mealli et al., 2004). These assumptions are problematic if, for example, people who are offered the treatment and who declined changed their attitudes toward the study itself compared to those who were not offered the treatment.

treat compliance as unobserved in the control state ignore any information, however, on the compliance behavior of controls that is potentially available through the study.

The statistical model we propose, the generalized endogenous treatment model (or GET model), uses a parametric item response model (e.g., Trier and Jackman, 2008) to measure the latent compliance type for all subjects, including those assigned to the control group. In contrast to the FR approach to principal stratification, GET measures compliance based on observed behavior rather than model-based imputation that relies heavily on the explanatory power of covariates. Because it incorporates additional information regarding compliance type, for both those assigned to the treatment and those assigned to the control, GET has the potential to improve the accuracy and efficiency of treatment effect estimates. In addition, GET retrieves the correlations among subjects' probabilities of taking up the treatment, probability of responding on an outcome measurement, and the outcome dependent variables, allowing a test of whether dependence remains even after balancing on observed covariates.

2.1. *Structural Parameters*

GET is a generalization of the random effect models proposed in Aakvik et al. (2005), Miranda and Rabe-Hesketh (2006), and Terza (1998); in particular, it extends the random effect approach to identifying treatment effects found in Aakvik et al. (2005) to account for nonresponse. Define the experiment \mathbf{e} as a study that randomly assigns subjects to treatment and control conditions in an effort intended to estimate the average treatment effect on a set of outcomes, \mathbf{o} . 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 as follows (individual indexes are suppressed).

The exogenous vector \mathbf{z} indicates subjects' randomized group assignment; for simplicity, assume a binary treatment and no interference among units (SUTVA), so \mathbf{z} con-

tains only a scalar Z that is one if assigned to the treatment, zero otherwise. There are K separate (but potentially dependent) experimental outcomes in \mathbf{o} . For each subject, we only observe the outcome corresponding to her actual treatment status, $O_k \equiv ZO_{1k} + (1 - Z)O_{0k}$. Define the treatment outcome vector \mathbf{o} of length K to be the potential outcomes that are observed for a given subject, that is, conditional on the treatment she actually received. If the subject did not respond on the outcome measurement, her outcome is missing. For example, in the application below, the treatment outcome vector \mathbf{o} is observed if and only if the subject responded to a follow up survey administered at the completion of the study.

A vector \mathbf{c} of length M contains 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 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, $\mathbf{c}_{\sim \text{tr}}$, measuring compliance in the experiment. 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 other elements of $\mathbf{c}_{\sim \text{tr}}$ one could also build in separate compliance tasks into the experimental design, or one could use the subjects's past history of compliance that may be available through a survey firm's panel records.

To simplify the modeling task, assume the treatment is only available to those assigned to the treatment, $Z = 1$; that is, monotonicity is ensured by the research design. To evaluate the causal effect of the treatment in this context, define a treatment exposure indicator $T = (ZC_t)$, where \cdot is the indicator function. If available, pretreatment variables may aid in identifying the marginal effect of a treatment. Define the vector \mathbf{o}_{pt} of length K to be the pretreatment values of the outcomes corresponding to those measured in \mathbf{o} , if such measures are available. 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 (Morgan and Winship, 2007, 71). The vector \mathbf{e} also contains a subvector \mathbf{x} of exogenous pre-treatment variables; \mathbf{x} does not include a constant.

2.2. Outcome Equations

To be consistent with the application below, assume the outcomes are dichotomous, $O_k \in \{0, 1\}$, and the data generating process (DGP) is 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 vector $\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \lambda_k)$ of structural parameters in each of K regressions:

$$O_k^* = \alpha_{0k} + \alpha_{1k}T + \alpha_{2k}O_{pt(k)} + \mathbf{x}'\boldsymbol{\beta}_k + \lambda_k\eta_1 + \epsilon_k \quad (1a)$$

$$O_k = \begin{cases} 1 & \text{if } O_k^* > 0, \\ 0 & \text{if } O_k^* \leq 0. \end{cases} \quad (1b)$$

where η_1 is an estimated measure of each subject's propensity to comply with the experiment, to be defined below.⁴ We model the DGP $p(O_k = 1|\cdot)$ using the probit inverse link function and distributional assumptions for η_1 and ϵ_k that we state in the measurement model section below.

2.3. Measurement Model and Identification of Structural Parameters

Identifying the structural parameters $\boldsymbol{\theta}_k$ requires accounting for the endogeneity of subjects' compliance with their assigned treatment and attrition through non-response since

⁴Note that this is the same specification as the outcome equation in Horiuchi et al. (2007), which specifies the linear index as $\alpha_h ZC_t + \beta_h(1 - Z)C_t$. 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.

these compliance processes can be correlated jointly with the outcome variables even after matching or conditioning on measured covariates. As in all principal stratification models, GET identifies the structural parameters by conditioning the model on subjects' latent compliance type. For simplicity, we assume that subjects in the control condition do not have access to the treatment. In the FR approach to principal stratification, this assumption 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 assumes that compliance type is measurable 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

$$C_m^* = \alpha_{0m} + \mathbf{x}'\boldsymbol{\beta}_m + \lambda_m\eta_1 + \epsilon_m \quad (2a)$$

$$C_m = \begin{cases} 1 & \text{if } C_m^* > 0, \\ 0 & \text{if } C_m^* \leq 0. \end{cases} \quad (2b)$$

with $m \in \{t, r, \}$. In the dichotomous case, we model $P(C_m = 1|.)$ using the probit inverse link, again giving distributional assumptions for ϵ_m in the next section.

Using equation set (2), GET uses the behavioral indicators \mathbf{c} to estimate a common latent factor, η_1 , along with a vector of factor coefficients, $\boldsymbol{\lambda}_m$, measuring subjects' latent tendency to comply with all aspects of the experiment. 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 $\eta_1, \epsilon_k, \epsilon_m$, we impose the standard assumptions in random effects models required for identification (Skrondal and Rabe-Hesketh, 2004). We assume $\eta_1 \sim N(0, 1)$; $\epsilon_m \sim N(0, 1)$, and $cov(\eta_1, \epsilon_m) = 0$ for all m ; $\epsilon_k \sim N(0, 1)$, $cov(\eta_1, \epsilon_k) = 0$

for all k ; and $cov(\epsilon_m, \epsilon_k) = 0$ for all m, k .⁵ In addition, one of the free parameters in $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_k\}$ must be set to 1 to scale the latent variable η_1 . Finally, 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 made conditionally independent by allowing it to load on η_1 .

Simultaneous to the estimation of the measurement model, GET includes the common factor η_1 as a latent control variable in the K outcome equations, with coefficient vector $\boldsymbol{\lambda}_k$. By including η_1 , the GET model in effect holds subjects' behavioral compliance "type" constant as a way to identify the structural parameters. With the assumption of latent ignorability, compliance with the treatment, C_t , response on outcome measures, C_r , and the outcomes, \mathbf{o} , are conditionally independent. That is, under the assumption of latent ignorability,

$$O_k \perp C_t, C_r, T \mid \eta_1, \mathbf{x} \quad (3)$$

Indeed, this form of conditional independence is a standard assumption in the item response theory literature (Patz and Junker, 1999; Trier and Jackman, 2008). One can see this in equation (1) since omitting η_1 in the k th outcome equation would make the combined error term $\eta_1 + \epsilon_k$ correlated with treatment compliance, since by construction η_1 is correlated with compliance, and this omission would bias all structural parameter estimates.

Assuming latent ignorability, 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} . That is, the GET model assumes the decisions to comply with the treatment and to respond on the outcomes are correlated with the outcomes. Since all outcome and compliance variables are assumed dichotomous, the

⁵We 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.

correlations between the compliance behaviors and the outcomes can be retrieved with:

$$\rho_{O_k, C_m} = \frac{\lambda_k \lambda_m}{\sqrt{(\lambda_k^2 + 1)(\lambda_m^2 + 1)}} \quad (4)$$

for all m, k . To retrieve the correlations among the compliance variables using this equation, substitute $m' \neq m$ for k , and to retrieve the correlations among the outcomes, substitute $k' \neq k$ 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).

The GET approach assumes that subjects' unobserved compliance "type" can be characterized by a latent variable (see Aakvik et al., 2005). 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 requires that the compliance processes measured in \mathbf{c} are not deterministic, so that some subjects with a low expectation 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 expectation do not receive the treatment.

2.4. *Treatment Effect Heterogeneity*

GET can test for heterogeneous treatment effects, where treatment effects differ across compliance types (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 functional form estimates both the main effect of compliance type on the outcome, and the interaction between the compliance type and the treatment she actually received. 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.

2.5. *Complex Data Structures*

As in any multilevel model, GET is flexible enough to account for complex data structures. For example, both the distribution of compliance type and the treatment effects can vary across sites of the experiment by specifying an additional level of the model that groups subjects within sites. The distribution of compliance types can vary across sites using a level three random intercept, and the treatment coefficient can vary across sites with a level three random coefficient. Note that the variances of the outcomes also can be estimated in GET when they are identified. With dichotomous responses, none of the variances are identified, so the variance equations of the outcome models are suppressed.

2.6. *Model Identification and Estimation*

We estimate the parameters of the $m + k$ equations simultaneously as a multilevel model using Bayesian MCMC methods with data augmentation to simulate the posterior distribution (Imbens, 1997). Here, the outcomes and compliance indicators are level one observations and subjects are level two observations. Assuming the form of conditional independence implied in latent ignorability, and noting that the prior distribution for η_1 has no free parameters, the posterior distribution is

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \eta_1) \propto p(\mathbf{o}|\eta_1, \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{x}, T, \mathbf{o}_{\text{pt}}) \times p(\mathbf{c}|\eta_1, \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{x}) \times p(\mathbf{o}_{\text{pt}}|\eta_1, \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{x}) \times p(\boldsymbol{\beta}) \times p(\boldsymbol{\lambda}) \quad (5)$$

We implement the models below in the MCMC software `WinBUGS`, which alternately imputes missing values in \mathbf{o} and \mathbf{c} using estimated parameter values, 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 $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, although in some instances assigning informative priors is theoretically justified and/or can aid convergence. In particular, informative priors on factor coefficients 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, $T = \iota(ZC_t)$, where ι is the indicator function, as an endogenous regressor in the outcome equations (see equation set 1). In addition, response on the outcome equations is modeled as endogenous selection. Miranda and Rabe-Hesketh (2006) and Terza (1998) demonstrate that no exclusion restrictions are necessary for theoretical identification of the endogenous regressors or selection. One must prove identification of a specific model, however, and proving identification of complex multilevel models analytically 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 dataset, 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, in particular if the compliance indicators are unreliable the model will have trouble converging.

2.7. Retrieving Treatment Effects

The ATE mean parameter estimates the average effect of the treatment for all subjects in the experiment. The ATE for an outcome equation k is,

$$\Delta_k^{ATE} = \int_{-\infty}^{+\infty} [\Phi(\alpha_{0k} + \alpha_{1k} + \lambda_k \eta_1) - \Phi(\alpha_{0k} + \lambda_k \eta_1)] \phi(\eta_1) d\eta_1 \quad (6)$$

This representation assumes that the covariates have been mean differenced, and one is interested in effect estimates for an average subject in the sample. Note that constraining η_1 to zero reduces the structural GET model to probit and Δ_k^{ATE} retrieves the treatment mean effect for that case (the integral evaluates to one times the difference in probabilities).

3. COMPARISON TO OTHER METHODS

In this section we compare the strengths and weaknesses of GET to existing approaches for principal stratification and to matching.

3.1. GET as an Extension of Principal Stratification

In the FR approach, the method of principal stratification can only observe 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 since 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 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. Since GET uses an item response model to estimate compliance type, it retrieves a full distribution of estimated compliance type for each subject, and these estimates may be substantively interesting in and of themselves. Finally, GET can make treatment effect heterogeneity a function of latent compliance type, allowing the effect of the treatment to vary between compliers and noncompliers.

Implementing GET does require, however, that these additional compliance data 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.

3.2. *Combining GET with Matching*

Non-parametric 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 only 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, while 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 dataset 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 ignorable after balancing on observed covariates. That is, GET can use equation (4) 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 reweighted dataset 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 reweighted dataset where treatment subjects are matched to control subjects (the control matched data). The overall ATE is $\pi(ATT) + (1 - \pi)(ATC)$, where π is the rate of compliance with the treatment.

4. RETRIEVING A BENCHMARK ATE IN A SIMULATION STUDY

We use simulation methods to compare the performance of GET relative to existing approaches to principal stratification⁶ when the processes of compliance and response depend on unobserved factors. For the simulation, we draw fifty independent samples of 1,000 observations each, composed of compliers and never takers. To create a control group in each sample, we randomly assign half of the observations to the treatment group with indicator variable Z equal to one for those assigned to treatment and zero otherwise. Denote a draw from the standard normal random distribution as N ; to reduce notational clutter, each instance of N indicates a fresh draw.

In each dataset the treatment has no effect, for compliers as well as for noncompliers, but instead would have an apparent effect in a naive model only through compliance and response patterns. Specifically, for each iteration, we first created a set of three

⁶For 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>.

correlated “observed” variables, a vector \mathbf{x} , with $X_1 = N$, $X_2 = 0.3X_1 + 0.7N$, and $X_3 = 0.2X_2 + 0.8N$, and an orthogonal “unobserved” random variable $\eta_1 = N$. We use the observed and unobserved variables to simulate four correlated endogenous choice variables: an outcome variable, O , and three compliance variables, compliance with the treatment, C_t , response on the outcome survey, C_r , and an additional indicator of compliance, $C_{\sim tr}$. We constructed C_t as a binomial response, with a probit inverse link function and linear index $0.002X_1 + 0.003X_2 + 0.004X_3 + 2\eta_1 + 0.5N$, setting C_t to missing if $Z = 0$. We constructed C_r , similarly with linear index $0.003X_1 + 0.004X_2 + 0.005X_3 + \eta_1 + 0.5N$ and $C_{\sim tr}$ with linear index $0.005X_1 + 0.004X_2 + 0.003X_3 + 2\eta_1 + 0.5N$. We constructed the outcome O as a binomial draw with linear index $0.006X_1 + 0.005X_2 + 0.004X_3 + 2\eta_1 + 0.5N$, setting O to missing if $C_r = 0$. Notice that the treatment compliance indicator, $T = \iota(ZC_t)$, where ι is the indicator function, is not in the true outcome equation. Instead, T by construction is strongly correlated with the unobserved variable η_1 by way of C_t . As a consequence, the naive probit estimator, regressing the outcome on T and \mathbf{x} with η_1 omitted, will retrieve a positive and significant, but spurious, treatment effect.

We estimated the structural parameters with GET, FR, and matching using these 50 datasets, including the vector \mathbf{x} and T in the outcome equation but not η_1 . The results of these trials are diagrammed in figure 1. The first column of histograms compares point estimates of the average treatment effect rescaled into percent changes. Since the treatment effect is zero ($ATE=CACE=0$), these diagrams also graph bias in the estimates across samples. Both GET and FR are unbiased across samples. The average ATE across samples for GET is -1.2 percent and for FR is -2.5 percent. Notice, however, that there is more density in the tails of the histogram for FR compared to GET. This difference in efficiency is demonstrated in the second column of the diagram, where the mean squared error for GET is 46.2 while the mean squared error for FR is 67.7, which is nearly 50 percent larger than GET. The final row of the table gives the matching results, which by the design of the dataset should be biased as the main variable driving compliance and

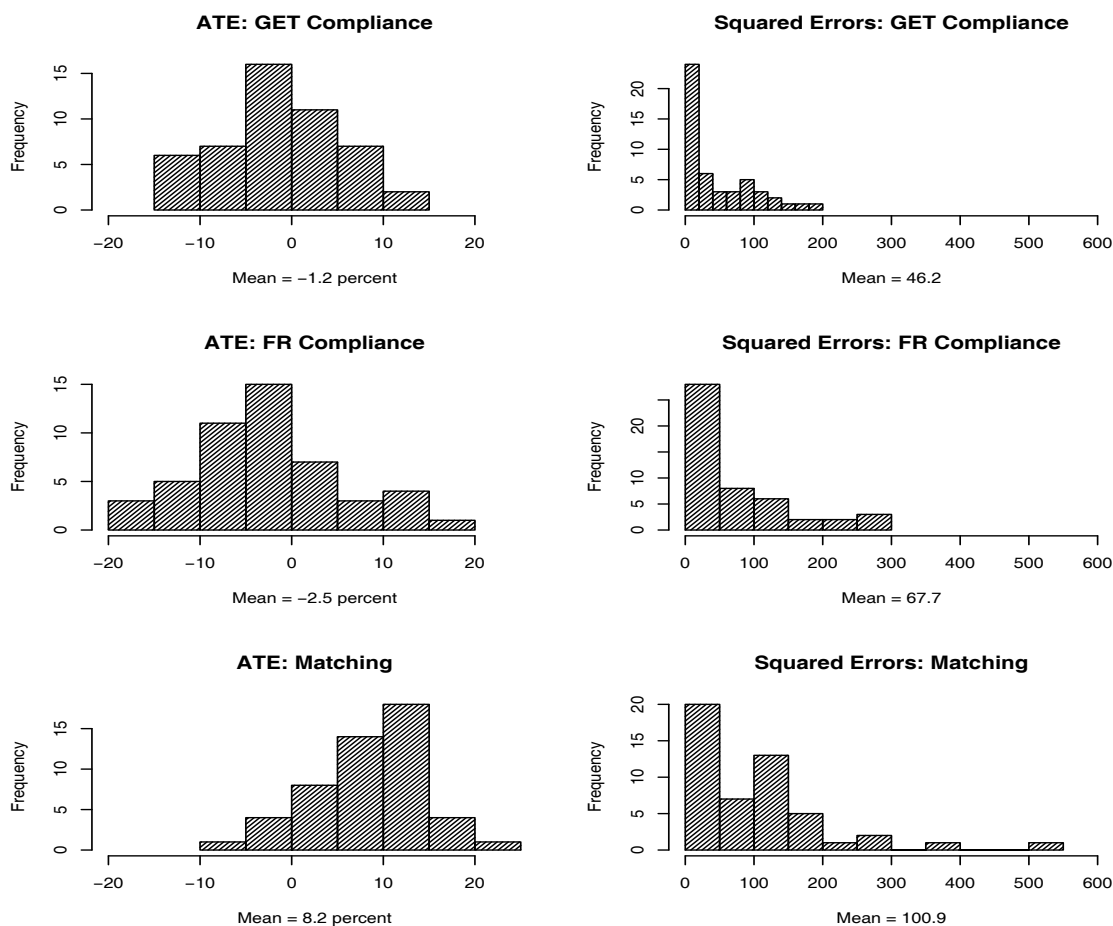


Figure 1: Comparison of ATE Estimates in the Monte Carlo Study

response, η_1 , is unobserved. As expected, matching retrieves a large and positive, but spurious, average treatment effect estimate, and as a consequence also has a high mean squared error.

5. APPLICATION: CITIZEN EFFICACY IN A DELIBERATIVE EXPERIMENT WITH MEMBERS OF CONGRESS

In this section, we show how to apply GET in practice, and compare GET to matching in a case where we have strong *a priori* expectations that compliance is driven by unobservables. 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 conducted either one or two sessions. In addition, we conducted two sessions substituting a noted policy expert on immigration for the member. With the two expert sessions, there were a total of twenty-two sessions. The number of participants in each session ranged from eight to thirty 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 thirty-five minutes, 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 Foundation⁷ 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), a respected online survey research firm, recruited the constituents from each congressional district and administered the surveys.⁸ After completing a pretest survey, each was randomly assigned to one of three conditions: a deliberative

⁷CMF is a nonpartisan, nonprofit, dedicated to assisting Members of Congress to better manage their offices. See www.cmfweb.org.

⁸In 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.

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.⁹ A week after each session, KN administered a follow up survey to subjects in each of the groups. That is, all subjects in a given congressional district received the follow up survey at the same time irrespective of the treatment actually received.

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 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.

5.1. *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

⁹In addition, the pretest informed subjects of the specific date and time of the session for their district, and asked if they were willing and able to participate. In the analysis below, we exclude those who responded no to this filter question.

survey, all subjects were asked “Please tell us how much you agree or disagree with the following statements:”

- I don’t think public officials care much what people like me think.
- 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 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 non-random treatment compliance and non-response 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 con-

trol groups for these outcomes, since subjects are unlikely to select themselves into the treatment anticipating disproportionate gains in efficacy.¹⁰

5.2. 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 pre-treatment responses from the baseline survey ($N = 670$) as well as post-treatment responses from the follow up survey ($N = 527$ with 143 non-responses). On both surveys, small percentages of subjects disagreed with the statement that officials do not care what ordinary people think, with only 15 percent disagreeing on the baseline survey and 22 percent 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 percent agreeing on the baseline and 68 percent 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 post tests (see Morgan and Winship, 2007, 71). As a consequence, the analyzes below treat the baseline outcomes as endogenous variables.¹¹

We have three measures of the propensity to comply with the study, corresponding to the indicator variables $C_t, C_r, C_{\sim tr}$. First, note that attending a session requires a fairly large commitment, including setting aside an hour to participate, and to do so at

¹⁰Heterogeneous 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 benefit from the program compared to those who choose not to enroll, leading to heterogeneous treatment effects.

¹¹Morgan and Winship (2007, 71) demonstrate that matching or conditioning on a pre-treatment outcome measure will typically bias treatment effect estimates, since the pre-treatment and post-treatment outcome measures are typically co-determined. In this application, the outcomes are measured in a survey, and so the pre- and post-treatment responses have stochastic components (see Achen, 1975) that will be correlated. Below we show how to include baseline outcome measures as endogenous regressors in GET.

a preset date and time. Because of this, we anticipated a low response rate, and initially assigned 83 percent of this sample to the deliberation arm, of which 53 percent actually *Participated in a deliberative session*, or C_t . For those assigned to the control (information only) condition, C_t is missing. All subjects were administered the follow up survey, and 79 percent *Responded on the follow up survey*, or C_r . This response rate is high, but not unusually high since we have restricted the sample to those who responded to both the baseline survey and the background materials survey. We also administered a survey after the November elections to subjects who completed the follow up survey, and among those, 88 percent *Completed the November survey*, or $C_{\sim tr}$. $C_{\sim tr}$ is missing if $C_r = 0$. 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 four variables in the pre-treatment survey that we believed would be the best observed measures of subjects' latent propensity to participate in the online deliberative session. Two of these variables 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 percent 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 percent liking this sort of responsibility.

Two other variables measure subjects' need for judgment (Bizer, Krosnick, Holbrook, Wheeler, Rucker, and Petty, 2004). On the pre-treatment 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 percent 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 percent reporting a preference not to be neutral about complex issues.

We have a number of other exogenous variables measuring various attributes of subjects that potentially can co-determine compliance with the treatment, outcomes, and response on the outcomes: 46 percent of subjects were *Not employed*; 39 percent had *Completed some college*; 45 percent had *Completed college or more*; 66 percent were *Female*; 84 percent *White*; 10 percent were in a *March (expert) session*; 50 percent came from a representative *KN panel* as opposed to an opt-in panel that KN subcontracted with; 68 percent were able to answer two or more items on a pre-test of the "Delli Carpini and Keeter five" items (Deli Caprini and Keeter 1993) indicating *High political knowledge*. Together, we call these exogenous variables the "attribute variables" and condition the model on all of them.

5.3. *Propensity Score Estimation and Balance in Matched Datasets*

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 judgment 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, Imai, King, and Stuart, 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 also included the exogenous variables listed above (minus the congressional district fixed effects), discarding all subjects outside of the support of the propensity score.

We used `MatchIt` and `GenMatch` to create two preprocessed datasets. For the first dataset we used the naturally coded treatment indicator (1 if treatment, 0 if control) to optimally match controls to treatments. In this dataset, 341 of the original 373 controls were matched (with 2 treated and 1 control discarded). For the second dataset, we used the reverse coding of the treatment indicator (1 if control, 0 if treatment) to optimally match treatment subjects to controls. In this dataset, 181 of the original 297 treatment subjects were matched. A comparison between treatment and control in the first dataset, the “treatment matched” dataset, identifies the average treatment effect for the treated (ATT), and those between control and treatment in the second dataset, the “control matched” dataset, identifies the average treatment effect for the controls (ATC). A weighted average of these two estimands yields the average treatment effect (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 dataset, 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 non-zero weight only to observations that are retained as matches. In the treatment matched dataset, 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 dataset, 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 next, we use an estimation approach (a Bayesian model implemented in `WinBUGS`) that does not allow weights. As a consequence, to preserve balance, we created two weighted-resampled matched datasets, one for the treatment matches and one for the control matches. The appendix describes how we resampled to create the weighted matched data (from here on we will refer to these simply as the matched data). Table 1 reports the descriptive statistics of each matched dataset, as well as the univariate balance scores for each exogenous variable. The balance score is calculated as the difference in the matched average between treatment and control groups, divided by the square root of the sum of the within group variance in the full data set. The maximum imbalance in the treatment matched dataset is -0.13 standard deviations, and the maximum imbalance in the control matched dataset is 0.33 standard deviations. Overall, the balance in both matched datasets is quite high, reflecting the strong performance of `GenMatch`.

5.4. *Estimation*

We estimate the structural parameters for the GET model using Bayesian MCMC as implemented in `WinBUGS` (Spiegelhalter, Thomas, Best, and Gilks, 1996), calling `WinBUGS` from the R software `R2WinBUGS`.¹² To identify the scale of the latent compliance variable we set the structural coefficient (λ_k) in the have ideas about politics outcome equation equal to one. We specified the priors for the structural coefficients in two ways. First, we

¹²<http://www.stat.columbia.edu/gelman/bugsR/>

estimated the model using independent normal diffuse priors for all parameters, including the factor coefficients, each with mean 0 and variance 1,000. Second, we estimated the model using these diffuse normal priors for all coefficients except for the factor coefficients, which instead were given informative uniform priors on $[0,4]$. The latter informative priors are useful to examine for two reasons. First, substantively, we have strong prior beliefs that all of the compliance and outcome efficacy indicators have positive correlation. Second, while the scale of the latent variable is set in the have ideas about politics outcome equation, in practice the MCMC model can have trouble converging, given the complexity of the simultaneous problems of imputing missing data and estimating the latent variable parameter distribution for each subject. In this case, the informative priors aid convergence considerably.

We apply the GET model to both the full data set and to the two matched datasets using each set of priors. In each case, we estimate three chains using randomized starting values. In the case of non-informative priors, the model will only converge if the majority of the initial values have the “correct” sign, so we assign these log normal priors with mean 0 and variance of 0.5. Notice that assigning the initial values in this way does not constrain the sign of the posterior point estimates. The chains in the GET model typically converge after a run of 10,000 iterations. We find that assigning informative priors does not change any treatment effect estimates within two decimal places compared to diffuse priors, save for one estimate: with diffuse priors, the ATE for the officials care outcome and its standard error in the full dataset were double that found with the informative priors. That is, out of six ATE estimates, only one estimate was affected by restricting the factor coefficients to be non-negative, and even then the estimate retained an equal ratio to its standard error. In addition, the models with informative priors were quicker to converge, converged to identical results in multiple trials, and had more symmetric marginal posterior distributions compared to the model with uninformative priors. For these reasons, we report below the results of the models with informative priors. As a

significance test for the factor coefficients bounded at zero, we report whether the 2.5 percent quantile of the marginal posterior distribution exceeds zero.

For the results we present below, we draw 50,000 iterations, discard the first 25,000 draws, and retain one in every twenty five draws for a total of 1,000 draws to simulate the posterior distribution. We use the simulated marginal posterior distributions to estimate the standard errors of all parameters, including functions of the structural parameters such as the latent correlations and the mean treatment effect parameters.

5.5. Results

Table 2 provides the results for the mean treatment effect estimates. The table is partitioned into four quadrants. The top portion of the table shows the mean treatment effect estimates from the probit model, and the bottom portion shows the mean treatment effect estimates and latent correlations retrieved from the GET model. 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 datasets. In addition to the treatment indicator, each model conditions on the full list of attribute variables, as well as on the baseline pre-treatment outcome. For the matched datasets, the ATT, ATC and ATE estimates are reported separately for each model. We have highlighted in bold the basic results that we wish to focus on in comparing across datasets and across estimators.

The naive probit model, or the probit model applied to the full data set, estimates large and precise estimates of the average treatment effect 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 percent ($p < 0.05$), and that they themselves have ideas about politics and policy that government officials should listen to (internal efficacy) by 15

Table 2: Treatment Effect and Latent Correlation Estimates

	Full Data Set		Matched Data Set	
	Officials Care (Post Treatment) Δ^{ATE}	Have Ideas About Politics (Post Treatment) Δ^{ATE}	Officials Care (Post Treatment) Δ^{ATE}	Have Ideas About Politics (Post Treatment) Δ^{ATE}
Probit Model				
ATE	0.087*	0.154*	0.096*	0.133*
ATT			0.143*	0.124*
ATC			0.043*	0.142*
GET Model				
ATE	0.069*	0.009	0.078*	-0.002
ATT			0.123*	-0.007
ATC			0.027*	0.003
Latent Correlations[†]				
Participated in Session	0.084	0.420*	0.103*	0.404*
Resp. Follow Up Survey	0.140	0.699*	0.177*	0.696*
Resp. November Survey	0.034	0.175*	0.051	0.211*
Officials Care (Pre-treat.)	0.025	0.125*	0.019	0.068*
Have Ideas (Pre-treat.)	0.007	0.038	0.008	0.032

* $p < 0.05$. [†]Latent correlations in the matched data represent weighted averages for treatment and control, and correspond to the ATE estimates.

Notes: Coefficients are the mean of the simulated posterior distribution draws, and standard errors are the standard deviations of the simulated draws. Statistical significance for the coefficients is determined from the 2.5 percent and 97.5 percent quantiles of the simulated distributions. Coefficients emphasized in the text are highlighted.

percent ($p < 0.05$). One would be incautious, however, to interpret these estimates as unbiased ATE estimates, since there is a strong chance that subjects with high efficacy are selecting into the treatment.

Recall that the matched datasets 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. Consider the latent correlation estimates between participating in the deliberative sessions and the outcome response variables that are retrieved in the GET model from the matched data. Even after achieving near perfect balance on the observables, there remains a correlation of 0.4 ($p < 0.001$) between participating in the deliberative session and the belief that one has ideas about politics and policy that people in government should listen to. That is, it appears that subjects are selecting into the deliberative sessions based on an unmeasured propensity that codetermines this measure of internal efficacy. 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 0.1 ($p < 0.05$). This suggests that subjects are selecting into the treatment based on internal more than on external efficacy.

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 and need for judgment. Indeed, heretofore, the empirical literature on deliberative democracy has yet to develop a highly predictive model of the propensity to participate in deliberation.

Once the GET model accounts for the latent dependence between participating in a session and internal efficacy, the deliberative sessions appear to have little effect on subjects' levels of internal efficacy; the ATE in this equation is only -0.002 (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.

In contrast, given there does not appear to be dependence between complying with the treatment and the measure of external efficacy, the GET model and the naive model 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 percent ($p < 0.05$). This estimate is statistically equal to the naive probit estimate, and by both models the analyst would be led to reject the null hypothesis.

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, while others may come to realize that their ideas were perhaps poorly informed compared to other discussion participants' statements. 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, should 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 results of the GET model as applied to the matched data are doubly robust, in that they account for both compliance with the treatment based on observed covariates and unobserved latent traits that are correlated with internal efficacy. As such, we will use these results to evaluate the performance of two other estimators, the probit estimator using matched data, and the GET estimator using the full data set. In addition

to being useful when combined with matching, the GET model may prove useful as a stand-alone alternative to matching in some situations. As Ho et al. (2007, 216) note, “The main diagnostic of success in matching is...balance, as well as the number of observations remaining after matching.” Since the GET model does not require that we discard observations, it can handle situations in which matching seems to result in an unduly desiccated data set.

If matching on observed covariates accounted for the dependence between compliance with the treatment and the outcomes, then the probit estimator applied to the matched data should identify the ATE. Notice that while the probit estimator does retrieve the correct ATE estimate for the effect on external efficacy, 10 percent ($p < 0.05$), it dramatically overestimates the ATE for internal efficacy. Indeed, the probit ATE estimate for the matched data is nearly identical to the naive estimate, substantively and statistically, which again reinforces the assertion that the observed measures available for matching cannot account for the propensity to comply with a deliberative experiment.

In contrast, the GET model does a reasonable job of retrieving both ATEs from the full data set. The point estimate for the internal efficacy ATE with the full data set is nearly identical to the internal efficacy ATE estimated using GET given 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. The GET point estimate for the external efficacy ATE based on the full data set also is identical to that from the matched data, and again the analyst would be led to make the same decision in both datasets to reject the null hypothesis. Overall, the GET model estimates remains robust across the two datasets since GET is accounting for the latent dependencies between treatment compliance, response, and outcome, while matching cannot account for these unobserved dependencies.

The ATT and ATC are identified for all estimators in the matched data and for a specific functional form in the GET model using the full data set. Recall that matching

creates two datasets, one where the controls are matched to the treatment subjects, and one where the treatment subjects are matched to the controls. Under the matching assumptions, estimators applied to the former identify the ATT, while those applied to the latter identify the ATC, (and a weighted average of the two estimands identifies the ATE). Both the GET and probit estimators applied to the matched data show constant treatment effects for the measure of internal efficacy, in that those who select into the treatment have about the same benefit from participating as those who do not select into the treatment. This suggests subjects are not selecting into the sessions in anticipation of a gain in internal efficacy. In contrast, the GET estimator shows some treatment effect heterogeneity on the external efficacy response. This suggests that perhaps those who selected into the sessions were the ones most open to the possibility that government officials will care about their thoughts on policy.

The GET model can retrieve heterogeneous treatment effects when the heterogeneity is a linear function of the latent propensity to comply with the study. To estimate this effect, one simply interacts the latent variable η_1 with the treatment indicator, where the main effect of the treatment identifies the average treatment effect for a subject with the mean propensity to comply, and the interaction term is multiplied by a random coefficient (see Aakvik et al., 2005). For example, if the coefficient on the interaction term were positive, then subjects with a high propensity to take up the treatment would have a higher response to the treatment, and vice versa for a negative estimated coefficient. In this application, we discovered no heterogeneity in this functional form using the full data set (results not reported). This means that the heterogeneity observed in the matched data for external efficacy is along some dimension other than the (estimated) latent propensity to comply with the study.

Turning attention to the latent correlations among the other endogenous variables, notice that the GET model retrieves similar latent dependencies among all endogenous variables whether using the matched or the full data set. The correlations reported in

table 2 are a function of the factor coefficient estimates reported in table 3, retrieved from equation 4. That the correlation estimates are stable across these datasets reinforces the proposition that much of the dependence in these data is driven by unobservables, and that GET can robustly identify and account for these dependencies even with the unmatched data. The GET model identifies the latent propensity to comply by allowing multiple behavioral measures of compliance to load on the latent variable, including the choice to attend a deliberative session if assigned, the choice to respond to the follow up survey, and the choice to respond to the November post election survey. All of these variables are correlated with the outcome. Note in particular that the GET model's estimates account for the correlation between the outcome response and the choice to respond to the follow up survey. This means that the GET model accounts for endogenous non-response in addition to non-random non-compliance. Finally, correlations among the indicators retrieved from equation 4 show the validity of the latent compliance measure. The correlation between participating in a discussion and responding on the follow up survey is 0.59; 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 0.24, all with $p < 0.05$.

The GET model also can account for dependence between endogenous regressors (other than the endogenous treatment) and the outcome responses. We note that pretreatment measures of survey outcomes tend to be codetermined with the post treatment measures. 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 statistically significant correlation between pretreatment and post treatment responses on the have ideas about politics measure of internal efficacy, but the level of correlation is substantively small. No pre-post correlation is observed for the officials care measure of external efficacy. As a consequence, we are able to justify including the baseline responses as exogenous control variables in the probit models.

Table 3: Factor Coefficient Estimates

	Full Data Set		Matched Data Sets			
			Match to Treatment		Match to Control	
	λ_k	S.E.	λ_k	S.E.	λ_k	S.E.
Officials Care						
Post-treatment	0.1	0.2	0.1	0.1	0.2	0.2
Pre-treatment	0.2	0.1	0.1	0.1	0.1	0.1
Have Ideas about Politics						
Post-treatment	1	—	1	—	1	—
Pre-treatment	0.1	0.1	0.0	0.1	0.1	0.1
Participated in Session	0.7*	0.3	0.7*	0.1	0.7*	0.1
Resp. Follow Up Survey	2.4*	0.6	1.9*	0.4	2.2*	0.5
Resp. November Survey	0.3	0.1	0.4	0.2	0.2	0.2

* $p < 0.05$.

Note: Cells give factor coefficients and standard errors for the GET model applied to each dataset.

6. DISCUSSION

We use the insight of Ho et al. (2007) that parametric estimators may be applied to matched datasets, to show the performance of the parametric approach in GET compared to nonparametric matching estimators. We show that GET provides a useful diagnostic to test for the ignorability of treatment compliance after matching. 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 judgment, 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. This selection process is plausible, suggesting that people choose to select into 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.

Given the GET model applied to the matched data yield doubly robust estimates, we find that there is a strong positive treatment effect for external efficacy but not internal efficacy. The results regarding internal efficacy are sensible since some subjects observe

that their Member or fellow citizens have better ideas than they have, and others will observe the opposite. Similarly, the results regarding external efficacy are sensible since subjects are more likely to believe that government officials care about what they think after participating in a deliberative session when they are able to observe first hand their member of Congress taking their comments and questions seriously.

When compared to these doubly robust findings, we find that matching alone identifies the treatment effect for external efficacy, which is not strongly correlated with compliance type, but tends to overestimate the treatment effect for internal efficacy. This further demonstrates that matching alone cannot compensate for the disproportionate selection of highly (internally) efficacious subjects into the treatment. In contrast, we find that the GET model applied to the full dataset tends to retrieve similar estimates to the doubly robust findings, which further suggests that much of the selection can be accounted for by unobservables. The application also shows how GET can account for non-random non-response, as the model assumes that the outcomes are correlated with the decision to respond on the final survey. Finally, the application shows how GET can accommodate endogenous variables, such as pretreatment outcome responses.

7. CONCLUSION

In any application where compliance and response are self-selected, there is likely a strong danger that selection into and out of the study are 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 identifying 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, such as response history data often collected by survey firms that use panels. 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.

A. APPENDIX ON CREATING WEIGHTED-RESAMPLED MATCHED DATASETS

When using estimators that do not take weights, one must create simulated balanced datasets, where the simulation selects observations based on their weights assigned in a matching algorithm. To simulate the weighted matched dataset for the treatment subjects, we first retained all treatment subjects. We then sorted the matched controls by their weights, and created a variable that partitioned a line segment of 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 dataset. In both cases, we created a series of ten simulated datasets. Then among these datasets, we retained the dataset that had the best balance, given the natural sampling variation inherent in the simulation. We retain the best balanced dataset, rather than a random dataset, since

standard practice is to retry matching until a balanced dataset is produced (Ho et al. 2007). An alternative would be to re-estimate each of the models below with all ten datasets and then averaging the distributions of results.

REFERENCES

- Aakvik, A., J. J. Heckman, and E. J. Vytlačil (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to norwegian vocational rehabilitation programs. *Journal of Econometrics* 125(March), 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), 1–18.
- Achen, C. H. (1975). Mass political attitudes and the survey response. *American Political Science Review* 69(Dec.), 1218–1231.
- 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(Nov.), 1062–1084.
- 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(June), 444–455.
- 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(462), 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(Oct.), 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(Feb.), 42–49.
- Cacioppo, J. T., R. E. Petty, and C. F. Kao (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment* 48(May), 306–307.
- Diamond, A. and J. S. Sekhon (2007). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. UC–Berkeley, Department of Political Science typescript.
- Frangakis, C. E. and D. B. Rubin (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86(2), 365–379.
- Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics* 58(March), 21–29.
- Ho, D. E., K. Imai, G. King, and E. A. Stuart (2004). Matchit: Matching as non-parametric preprocessing for causal inference. Technical report, Harvard University. <http://gking.harvard.edu/matchit/>.
- 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(Summer), 199–236.
- 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(July), 669–687.
- Imai, K. (2007, December). Statistical analysis of randomized experiments with nonignorable missing binary outcomes: An application to a voting experiment. Princeton University, Department of Government typescript. <http://imai.princeton.edu>.

- Imbens, Guido W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* 25(1), 305–327.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 86(Feb.), 4–29.
- Jackman, S. (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science* 44(April), 369–398.
- 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(2), 207–222.
- 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(3), 285–308.
- Morgan, S. L. and C. Winship (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, N.Y.: Cambridge University Press.
- Morrell, M. (2005). Deliberation, democratic decision-making and internal political efficacy. *Political Behavior* 27(March), 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(Winter), 342–366.
- 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(Feb.), 33–38.
- Rubin, D. B. (1974). Estimating casual effects of treatemtns in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multi-level, Longitudinal, and Structural Equation Models*. Boca Raton, Fla.: Chapman and Hall.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (1996). Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). Technical report, MRC Biostatistics Unit.

Terza, J. V. (1998). Estimating count data with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics* 84(May), 129–154.

Trier, S. and S. Jackman (2008). Democracy as a latent variable. *American Journal of Political Science* 52(Jan.), 201–217.

Young, I. M. (1990). *Justice and the Politics of Difference*. Princeton, N.J.: Princeton University Press.