Advance Access publication May 10, 2010

Political Analysis (2010) 18:297–315
doi:10.1093/pan/mpq008

Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments

Alan S. Gerber
Institution for Social and Policy Studies and Department of Political Science, Yale University, 77 Prospect Street, New Haven, CT 06511
e-mail: alan.gerber@yale.edu (corresponding author)

Donald P. Green
Institution for Social and Policy Studies and Department of Political Science, Yale University, 77 Prospect Street, New Haven, CT 06511

Edward H. Kaplan
School of Management, School of Public Health, and School of Engineering and Applied Science, Yale University, 135 Prospect Street, New Haven, CT 06511

Holger L. Kern
Institution for Social and Policy Studies, Yale University, 77 Prospect Street, New Haven, CT 06511. From August 2010, Department of Political Science, University of South Carolina, 817 Henderson Street, Columbia, SC 29208

Randomized experiments commonly compare subjects receiving a treatment to subjects receiving a placebo. An alternative design, frequently used in field experimentation, compares subjects assigned to an untreated baseline group to subjects assigned to a treatment group, adjusting statistically for the fact that some members of the treatment group may fail to receive the treatment. This article shows the potential advantages of a three-group design (baseline, placebo, and treatment). We present a maximum likelihood estimator of the treatment effect for this three-group design and illustrate its use with a field experiment that gauges the effect of prerecorded phone calls on voter turnout. The three-group design offers efficiency advantages over two-group designs while at the same time guarding against unanticipated placebo effects (which would undermine the placebo-treatment comparison) and unexpectedly low rates of compliance with the treatment assignment (which would undermine the baseline-treatment comparison).

Authors' note: The authors are grateful to Mark Grebner, who conceived of the intervention described here and assisted in data collection, and to the Institution for Social and Policy Studies. We also thank the editors and anonymous reviewers, who provided very valuable comments. The experiment reported in this article was reviewed and approved by the Human Subjects Committee at Yale University. Supplementary materials for this article are available on the *Political Analysis* Web site.

1 Introduction

Administering a treatment to randomly assigned individuals presents a range of practical problems. Subjects may be unwilling to participate in a study or unavailable to receive the treatment when it is administered. More generally, they sometimes ''cross over'' from one experimental condition to the other and take the treatment when assigned to the control group or vice versa. In the end, the treatment to which subjects were randomly assigned may differ from the treatment that subjects actually receive. The problem of noncompliance with treatment assignment has received extensive attention in the statistical literature (e.g., Efron and Feldman 1991; Imbens and Angrist 1994; Angrist et al. 1996; Imbens and Rubin 1997; Frangakis and Rubin 2002; Cheng and Small 2006). One fundamental insight has been the value of the placebo-controlled experimental design (Boruch 1997, chap. 9; Nickerson 2005). In this design, subjects who agree to participate in a study and for whom the prospect of treatment is imminent are randomly assigned to receive either the treatment or the placebo. For example, Nickerson (2008) orchestrated a canvassing campaign in which registered voters were either encouraged to vote (the treatment) or to recycle (the placebo). In principle, the comparison of those who receive the treatment with those who receive the placebo minimizes problems of noncompliance and enables the researcher to draw unbiased and relatively efficient inferences about the causal effect of the treatment on the treated.

The use of such placebo-treatment designs, however, has several potential disadvantages. First, administering a placebo can be costly. Resources devoted to recruit subjects and secure their compliance must be divided between the treatment and the placebo arms of the experiment. Second, special care must be taken to ensure that the treatment and placebo are administered in ways that preserve the comparability of the groups. Finally, placebo-treatment designs do not allow researchers to detect any unexpected effects of the placebo itself.

An alternative design involves a comparison between a randomly assigned control group, which receives no intervention of any kind, and a randomly assigned treatment group, some fraction of which complies with the treatment assignment and is actually treated. For example, from a population of N individuals, one might select N_T subjects at random and attempt to treat them, leaving $N - N_T$ subjects in an untreated baseline group. Under certain assumptions that will be detailed below, this approach enables the researcher to consistently estimate the average treatment effect for those who actually receive the treatment. Whereas the placebo-treatment design compares those who receive the treatment with those who receive the placebo, the baseline-treatment design compares those assigned to the treatment group with those assigned to the baseline group, adjusting for noncompliance. Useful estimates may be generated using either approach. But the baseline-treatment design, too, has risks. If the compliance rate is low, the experiment may lack the power to detect a treatment effect. In extreme cases, compliance may be so low as to lead to substantial finite sample bias even with huge samples (Bound et al. 1995; Imbens and Rosenbaum 2005; Morgan and Winship 2007, 197–200; Angrist and Pischke 2009, 205–216).

The aim of this article is to show how these two designs can be combined to produce more accurate treatment effect estimates. We consider the case in which subjects are randomly assigned to three groups—an untreated baseline group, a placebo group, and a treatment group (Rosenthal 1985). After explicating the identifying assumptions underlying this approach using the Rubin Causal Model (Rubin 1974, 1977, 1978; Holland 1986), we derive a maximum likelihood (ML) estimator for the three-group design. Prior

statistical work has analyzed the relative merits of the placebo-treatment and baseline-treatment designs but has not discussed the advantages of the three-group design over these more conventional two-group designs.[1]

Of special interest is the advantage of the three-group design over the placebo-treatment design. There are many situations in which researchers conduct placebo-controlled experiments using samples that are drawn from large populations for which experimental outcomes can be measured at little or no additional cost. Examples include field experiments in which outcome data, such as voter turnout, mortgage loans, campaign contributions, crime rates, or mortality, are routinely supplied at low cost by public agencies. A researcher studying these outcomes using a placebo-treatment design would find it virtually costless to measure outcomes for an untreated baseline group as well. The purpose of this article is to derive an estimator for the three-group design and to lay out the conditions under which it provides improved statistical precision.

The article is organized as follows. We begin by deriving the three-group estimator within the potential outcomes framework for causal inference, building on previous work on randomized experiments with noncompliance (e.g., Imbens and Angrist 1994; Angrist et al. 1996). Next, we show how to implement the estimator using ML. We then illustrate its use with a field experiment. The study, which tests whether prerecorded phone calls increase voter turnout, shows the advantages of the three-group design in terms of added precision. We also present a small simulation that illuminates the conditions under which researchers might want to use a three-group design instead of the placebo-treatment design. These results also address the question of optimal allocation of resources when baseline observations are free and placebo and treatment observations are equally costly. We conclude by discussing other advantages of the three-group design beyond its increased precision, such as its ability to guard against unanticipated placebo effects (which would undermine the placebo-treatment comparison) and unexpectedly low rates of compliance with the treatment (which would undermine the baseline-treatment comparison).

2 Treatment Effect Estimation with Noncompliance

Suppose that we are interested in the effect of a binary treatment $D \in \{0, 1\}$ on some outcome Y. In the literature on voter mobilization, $D = 1$ represents an encouragement to participate in the upcoming election and Y is observed turnout in that election. As in Rubin (1974, 1977, 1978), we define Y_{1i} and Y_{0i} as the potential outcomes that individual i would have with and without being exposed to the treatment. In the voter mobilization example, Y_{1i} represents i's potential turnout after being encouraged to vote, whereas Y_{0i} represents i's potential turnout after not being encouraged to vote. The ''fundamental problem of causal inference'' (Holland 1986) is that we cannot observe both potential outcomes

[1]The problem considered here is similar but not identical to that encountered in three-arm trials with noncompliance (see, e.g., Cheng and Small 2006). Three-arm trials generally compare two active treatments to a control condition and to each other. Although the three-group design proposed here also has three treatment arms (baseline, placebo, and treatment), it invokes a different set of identifying assumptions. In classical three-arm trials, researchers are interested in estimating the effects of two treatments, whereas we are only interested in estimating the effect of a single treatment; the placebo group is solely used to allow for more efficient treatment effect estimates. Also note that there is a subtle yet important difference between the placebo effect considered here and placebo effects in clinical trials. In clinical trials, the placebo effect is the psychological effect of administering biologically inactive substances or procedures (see de Craen et al. 1999). Here, however, the placebo effect is the effect of administering a second treatment that we assume to have no effect *on the outcome of interest*, but which might very well have an effect on other outcomes. For example, an appeal to recycle, the placebo treatment used in Nickerson 2008, is assumed to not affect voter turnout but might very well affect recycling behavior and environmental awareness more generally.

Y_{1i} and Y_{0i}; we only observe $Y_i = D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}$. Since one of the two potential outcomes is always counterfactual, we cannot compute the individual-level treatment effect, $Y_{1i} - Y_{0i}$. Instead, we might want to estimate the average treatment effect, $\mathbb{E}[Y_1 - Y_0]$, or the average treatment effect on the treated, $\mathbb{E}[Y_1 - Y_0|D = 1]$.

Individuals do not always comply with their treatment assignment. Individuals assigned to receive a reminder to vote, for example, might not be contactable. Instrumental variable methods can be used to estimate treatment effects in randomized experiments with such noncompliance. Let $Z_i \in \{0, 1\}$ denote the treatment assigned to i and D_i the treatment actually received by i. Because of noncompliance the assigned treatment, Z_i, does not necessarily equal the received treatment, D_i. We follow Imbens and Angrist (1994) in conceptualizing the identification of treatment effects in terms of potential treatment indicators. Let D_{zi} represent i's potential treatment status given $Z_i = z$. For example, $D_{1i} = 1$ and $D_{0i} = 0$ means that i would take the treatment when assigned to it but would not take the treatment when not assigned to it. The treatment status indicator can then be expressed as $D_i = Z_i \cdot D_{1i} + (1 - Z_i) \cdot D_{0i}$. We only observe Z and D (and therefore D_z for individuals with $Z = z$) but never both potential treatment indicators for the same individual. Following the terminology in Angrist et al. (1996), we divide the population into four types defined by their potential treatment indicators D_1 and D_0: compliers ($D_0 = 0$ and $D_1 = 1$), always-takers ($D_1 = D_0 = 1$), never-takers ($D_1 = D_0 = 0$), and defiers ($D_0 = 1$ and $D_1 = 0$). Since we only observe one of the two potential treatment indicators, we cannot directly infer the type of any particular individual without imposing further restrictions.

Let Y_{zdi} represent the potential outcome that i would obtain if $Z_i = z$ and $D_i = d$. With a binary instrument Z_i and binary treatment D_i, the potential outcomes are Y_{00i}, Y_{01i}, Y_{10i}, and Y_{11i}.[2] In the voter mobilization example, Y_{10i}, for instance, represents i's potential turnout if she were assigned to receive a reminder but did not receive it.

Under the following assumptions, average treatment effects are nonparametrically identified for the subpopulation of compliers (Angrist et al. 1996; Abadie 2003):[3]

(i) Ignorability of the instrument: the random vector $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$ is independent of Z.

(ii) Exclusion of the instrument: $P(Y_{0d} = Y_{1d}) = 1$ for $d \in \{0, 1\}$.

(iii) First-stage effect: $P(D_1 = 1) > P(D_0 = 1)$.

(iv) Monotonicity: $P(D_1 \geq D_0) = 1$.

Assumption (i) is automatically satisfied in a randomized experiment because of random assignment of Z. Assumption (ii) asserts that variation in the instrument does not change potential outcomes other than through D and therefore allows us to define potential outcomes in terms of D alone: $Y_0 = Y_{00} = Y_{10}$ and $Y_1 = Y_{01} = Y_{11}$.[4] Together, assumptions (i) and (ii) guarantee that the only effect of the instrument on the outcome is through the variation it induces in the treatment status. The first-stage effect assumption (iii) requires that treatment assignment, Z, affects the actual treatment received, D. The strength of the observed relationship between Z and D is easy to assess empirically. Assumption (iv) rules

[2] Implicit in this notation is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1990), which requires that no individual is affected by the treatment assigned to and received by other individuals.

[3] We suppress the individual-specific index i to simplify the notation in this section.

[4] In the voter mobilization example, this exclusion restriction implies that turnout after not getting a reminder when being assigned to not get a reminder, Y_{00}, is the same as turnout after not getting a reminder when being assigned to get a reminder, Y_{10}. It also implies that turnout after getting a reminder when being assigned to not get a reminder, Y_{01}, is the same as turnout after getting a reminder when being assigned to get a reminder, Y_{11}.

out the existence of defiers and defines a partition of the population into always-takers, compliers, and never-takers.

Given these assumptions, the Wald estimator identifies average treatment effects for compliers, also called local average treatment effects by Imbens and Angrist (1994). Following Angrist et al. (1996, 445),

$$\tau_{\text{LATE}} = \frac{\text{cov}(Y,Z)}{\text{cov}(D,Z)} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1]-\mathbb{E}[D|Z=0]} = \mathbb{E}[Y_1-Y_0|D_1>D_0]. \quad (1)$$

With three treatments, $D \in \{0, 1, 2\}$, and three treatment assignments, $Z \in \{0, 1, 2\}$, the number of types increases to $3^3 = 27$ (Imbens 2007). Now, 0 corresponds to the baseline, 1 is the placebo, and 2 is the active (here: voter mobilization) treatment. In contrast to the binary instrument/binary treatment case considered before, the set of types T is now equal to $T \in \{(0, 0, 0), (0, 0, 1), (0, 0, 2), (0, 1, 0), (0, 1, 1), (0, 1, 2), \ldots, (1, 0, 0), \ldots, (2, 2, 2)\}$. In consequence, the characterization of compliers and noncompliers is no longer as straightforward. For example, $T = (0, 0, 0)$ is a never-taker who takes the baseline when assigned to the baseline ($D_0 = 0$), the placebo ($D_1 = 0$), or the treatment ($D_2 = 0$). $T = (0, 1, 0)$ is a partial complier who takes the baseline when assigned to the baseline ($D_0 = 0$) and takes the placebo when assigned to the placebo ($D_1 = 1$) but takes the baseline when assigned to the treatment ($D_2 = 0$).

We make the following six identifying assumptions. The first three assumptions parallel assumptions (i) to (iii) made before in the case of binary instruments and binary treatments. The last three assumptions are derived from the particular structure of the three-group design.

(a) Ignorability of the instrument: the random vector (Y_{00}, Y_{01}, Y_{02}, Y_{10}, Y_{11}, Y_{12}, Y_{20}, Y_{21}, Y_{22}, D_0, D_1, D_2) is independent of Z.

(b) Exclusion of the instrument: $P(Y_{0d} = Y_{1d} = Y_{2d}) = 1$ for $d \in \{0, 1, 2\}$.

(c) First stage: $P(D_z = z) > P(D_{\neg z} = z)$ for $z \in \{0, 1, 2\}$.

(d) No baseline crossover: $P(D_0 \neq 0) = 0$.

(e) Limited crossover: $P(D_1 \neq 2) = P(D_2 \neq 1) = 1$.

(f) Perfect blindness: $P(D_1 \neq 1|D_2 = 2) = P(D_2 \neq 2|D_1 = 1) = 0$.

Assumption (d) requires that all individuals assigned to the baseline group comply with their treatment assignment. This assumption is often plausible since individuals assigned to the baseline group normally cannot receive the placebo or the treatment. The ''no baseline crossover'' assumption rules out all types with $D_0 \neq 0$ such as $T = (2, 0, 0)$ or $T = (1, 1, 1)$ and reduces the number of possible types to 9. It would be violated if members of the baseline group had access to the placebo or the treatment, perhaps because of another experiment that made the placebo or the treatment available to these individuals.

Assumption (e) prohibits crossover between the placebo and the treatment groups. This assumption is also plausible in the context of many three-group experiments since individuals generally cannot receive the treatment when assigned to the placebo or the placebo when assigned to the treatment. The limited crossover assumption rules out $T \in \{(0, 2, 0), (0, 2, 1), (0, 2, 2), (0, 0, 1), (0, 1, 1)\}$. We are thus left with four types of individuals: $T \in \{(0, 0, 0), (0, 1, 0), (0, 0, 2), (0, 1, 2)\}$. $T = (0, 0, 0)$ characterizes individuals who always take the baseline. $T = (0, 1, 0)$ characterizes individuals who take the baseline when assigned to the baseline or the treatment but take the placebo when assigned to the placebo. $T = (0, 0, 2)$ characterizes individuals who take the baseline when assigned to the baseline or the placebo but take the treatment when assigned to the treatment. Finally, $T = (0, 1, 2)$ denotes individuals who always comply with their experimental assignment. They take the baseline

when assigned to the baseline, take the placebo when assigned to the placebo, and take the treatment when assigned to the treatment.

Assumption (f) asserts that compliance does not vary with treatment assignment. Perfect blindness (see Efron and Feldman 1991) rules out the existence of partial compliers who comply with their assignment when assigned to the placebo ($D_1 = 1$) but fail to comply with their assignment when assigned to the treatment ($D_2 \neq 2$) or vice versa ($D_1 \neq 1$ and $D_2 = 2$). In the voter mobilization field experiment we consider below, perfect blindness is satisfied because individuals do not know whether the phone call they are about to receive contains the placebo or the treatment. Since compliance is defined as simply answering the call, an individual who would comply with the placebo would also comply with the treatment and vice versa. Perfect blindness allows us to eliminate both partial compliers ($T = (0, 1, 0)$ and $T = (0, 0, 2)$). We are thus left with only two types of individuals: never-takers ($T = (0, 0, 0)$) and compliers ($T = (0, 1, 2)$).

The perfect blindness assumption can be assessed empirically. It implies that in expectation, the proportion of individuals who receive the placebo in the placebo group equals the proportion of individuals who receive the treatment in the treatment group. Moreover, it also implies that in expectation, the covariate distribution for individuals who receive the placebo is the same as the covariate distribution for individuals who receive the treatment. Whether the perfect blindness assumption is satisfied in practice depends on the way in which the experiment is designed and executed. In voter mobilization field experiments, for example, perfect blindness would be less plausible if callers employed different scripts on different days. Although such an experimental protocol might be easier to implement, it raises the question of whether compliance varies with treatment assignment. If there were systematic differences between the types of voters who respond to calls on particular days of the week, the perfect blindness assumption would be violated, resulting in potential bias in treatment effect estimates.

As shown in Table 1, it is straightforward to infer the population shares of never-takers and compliers given assumptions (a)–(f). Consider individuals with $(Z_i, D_i) = (0, 0)$. Such individuals can be either never-takers or compliers. We cannot infer the type of these individuals from the observed data. However, now consider individuals with $(Z_i, D_i) = (2, 0)$. Such individuals can only be never-takers. The same is true for individuals with $(Z_i, D_i) = (1, 0)$. Individuals with $(Z_i, D_i) = (1, 1)$ or $(Z_i, D_i) = (2, 2)$ in contrast must be compliers. We can derive the population shares of never-takers and compliers by considering the subpopulation with $Z_i = 2$. Within this subpopulation, we observe $D_i = 0$ only for never-takers and $D_i = 2$ only for compliers. Hence, the population share of never-takers is equal to $P(D_i = 0|Z_i = 2)$ and the population share of compliers is equal to $P(D_i = 2|Z_i = 2)$.[5]

Table 1 Type by observed variables

	Assigned treatment		
Received treatment	*Baseline ($Z_i = 0$)*	*Placebo ($Z_i = 1$)*	*Treatment ($Z_i = 2$)*
None ($D_i = 0$)	(0, 0, 0), (0, 1, 2)	(0, 0, 0)	(0, 0, 0)
Placebo ($D_i = 1$)		(0, 1, 2)	
Treatment ($D_i = 2$)			(0, 1, 2)

[5]Equivalently, one can derive the population shares of never-takers and compliers by considering the subpopulation with $Z_i = 1$. Within this subpopulation, we observe $D_i = 0$ only for never-takers and $D_i = 1$ only for compliers.

These population shares for never-takers and compliers allow us to infer outcome distributions for compliers under baseline, placebo, and treatment. We can directly infer the distribution of $Y_i(D_i = 2, T_i = (0, 1, 2))$, that is, the distribution of outcomes for compliers who receive the treatment, from the subpopulation with $(Z_i, D_i) = (2, 2)$ since all these individuals are known to be compliers. The same is true for the distribution of $Y_i(D_i = 1, T_i = (0, 1, 2))$, which we can infer from the subpopulation with $(Z_i, D_i) = (1, 1)$. We can also infer the distribution of $Y_i(D_i = 0, T_i = (0, 0, 0))$ from the subpopulation with $(Z_i, D_i) = (2, 0)$ since all these individuals are known to be never-takers. Finally, we use the distribution of $Y_i(Z_i = 0, D_i = 0)$. This is a mixture of the distribution $Y_i(D_i = 0, T_i = (0, 1, 2))$ for compliers and $Y_i(D_i = 0, T_i = (0, 0, 0))$ for never-takers with mixture probabilities equal to their population shares. Since we already inferred the population shares of never-takers and compliers as well as the distribution of $Y_i(D_i = 0, T_i = (0, 0, 0))$, we can obtain the distribution of $Y_i(D_i = 0, T_i = (0, 1, 2))$.

Average differences between pairs of these three outcome distributions can be interpreted as average treatment effects on the treated.[6] $\tau_{bp} = \mathbb{E}[Y_i(D_i = 1, T_i = (0, 1, 2)) - Y_i(D_i = 0, T_i = (0, 1, 2))]$ is the effect of the placebo for individuals who received the placebo instead of the baseline. $\tau_{bt} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 1, 2)) - Y_i(D_i = 0, T_i = (0, 1, 2))]$ is the effect of the treatment for individuals who received the treatment instead of the baseline. $\tau_{pt} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 1, 2)) - Y_i(D_i = 1, T_i = (0, 1, 2))]$ is the effect of the treatment for individuals who received the treatment instead of the placebo.

Both the placebo-treatment design and the three-group design rest on the assumption that the effect of the placebo on the outcome of interest is zero, that is, $\tau_{bp} = 0$. This assumption distinguishes the use of these designs in field research with noncompliance from the use of placebo-controlled experiments in other contexts such as research in the life sciences, where placebos are used to control for the psychological effects of therapeutic attention (Rosenthal 1985; de Craen et al. 1999; Torgerson and Torgerson 2008). If $\tau_{bp} \neq 0$, the placebo has an effect on the outcome and τ_{pt} and τ_{bt} differ.[7]

What happens if assumptions (a)–(e) hold but perfect blindness is implausible? Without perfect blindness, we cannot eliminate the partial compliers ($T_i = (0, 1, 0)$ and $T_i = (0, 0, 2)$) and are left with four types of individuals: compliers, never-takers, and two types of partial compliers. One of them complies with the placebo but not the treatment; the other complies with the treatment but not the placebo. Table 2 shows the result. Since each nonempty cell now contains a mixture of at least two types, we can no longer disentangle the mixture probabilities of the four types.

Even without perfect blindness, we can make some progress by analyzing the data as if individuals had only been assigned to the baseline or the treatment. In other words, we drop all individuals assigned to the placebo, so now $D_i \in \{0, 2\}$ and $Z_i \in \{0, 2\}$. This removes partial compliers who would only comply with the placebo but not the treatment ($T_i = (0, 1, 0)$) by turning them into never-takers ($T_i = (0, 0)$). It also removes partial compliers who would only comply with the treatment but not the placebo ($T_i = (0, 0, 2)$)

[6] Instrumental variable methods generally only identify average treatment effects for compliers (Angrist et al. 1996). Here, in the absence of types other than never-takers and compliers, complier average treatment effects can be interpreted as average treatment effects on the treated since every treated individual must be a complier.

[7] Researchers can test for the existence of a placebo effect by using the baseline-placebo comparison. Unless τ_{bp} is assumed to be zero (or a known constant), the three-group estimator will not provide any leverage in estimating the treatment effect beyond that afforded by the baseline-treatment estimator. *Ex ante,* the three-group estimator is subject to two sources of uncertainty, statistical uncertainty arising from sampling variability and modeling uncertainty associated with the assumed absence of a placebo effect as well as assumptions (a)–(f) (see Gerber et al. 2004).

Table 2 Type by observed variables without perfect blindness

	Assigned treatment		
Received treatment	*Baseline ($Z_i = 0$)*	*Placebo ($Z_i = 1$)*	*Treatment ($Z_i = 2$)*
None ($D_i = 0$)	(0, 0, 0), (0, 1, 2), (0, 1, 0), (0, 0, 2)	(0, 0, 0), (0, 0, 2)	(0, 0, 0), (0, 1, 0)
Placebo ($D_i = 1$)		(0, 1, 2), (0, 1, 0)	
Treatment ($D_i = 2$)			(0, 1, 2), (0, 0, 2)

by turning them into compliers ($T_i = (0, 2)$). We are thus left with never-takers and compliers and the average effect of the treatment on the treated is again identified: $\tau_{\text{bt}} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 2)) - Y_i(D_i = 0, T_i = (0, 2))]$. Alternatively, we can drop individuals assigned to the treatment, which yields $\tau_{\text{bp}} = \mathbb{E}[Y_i(D_i = 1, T_i = (0, 1)) - Y_i(D_i = 0, T_i = (0, 1))]$, the effect of the placebo for individuals who received the placebo. Even without the perfect blindness assumption, τ_{bp} and τ_{bt} are still identified. Not identified, however, is τ_{pt}, the average effect of receiving the treatment instead of the placebo for compliers. To identify average treatment effects on the treated in placebo-treatment experiments and three-group experiments, both of which rely on identification of τ_{pt}, perfect blindness has to hold. In what follows, we assume that perfect blindness holds and present empirical evidence that supports the validity of this assumption.

3 Estimators

We first derive estimators for τ_{pt} and τ_{bt} and then introduce a ML estimator for the effect of the treatment on the treated that achieves improved efficiency by combining these two estimators.

3.1 *Placebo-Treatment Comparison*

As shown above, $\tau_{\text{pt}} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 1, 2)) - Y_i(D_i = 1, T_i = (0, 1, 2))]$ is the effect of the treatment for individuals who received the treatment instead of the placebo. Given assumptions (a)–(f), we can directly identify these two outcome distributions; we therefore have the following:

$$\tau_{pt} = \mathbb{E}\left[Y_i\left(D_i = 2, T_i = \left(0, 1, 2\right)\right) - Y_i\left(D_i = 1, T_i = \left(0, 1, 2\right)\right)\right] \tag{2}$$

$$= \mathbb{E}[Y_i(Z_i = 2, D_i = 2)] - \mathbb{E}[Y_i(Z_i = 1, D_i = 1)], \tag{3}$$

with sample analog

$$\widehat{\tau_{\text{pt}}} = \frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 2, D_i = 2\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 2\}} - \frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 1, D_i = 1\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 1, D_i = 1\}}, \tag{4}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The first term on the right-hand side of equation (4) is the proportion of voters in the group receiving the treatment; the second term on the right-hand side of equation (4) is the proportion of voters in the group receiving the placebo.

3.2 *Baseline-Treatment Comparison*

As shown above, $\tau_{\text{bt}} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 1, 2)) - Y_i(D_i = 0, T_i = (0, 1, 2))]$ is the effect of the treatment for individuals who received the treatment instead of the baseline. We can directly observe the first outcome distribution; the second we can infer using the mixture probabilities. Recall that the population share of never-takers is equal to $P(D_i = 0|Z_i = 2)$ and the population share of compliers is equal to $P(D_i = 2|Z_i = 2)$. We can write the conditional distribution of $Y_i(Z_i = 0, D_i = 0)$ (the top left cell in Table 1) as a mixture of two distributions with known mixture probabilities:

$$\begin{aligned} Y_i(Z_i = 0, D_i = 0) = {} & Y_i(D_i = 0, T_i = (0,\ 1,\ 2)) \times P(D_i = 2|Z_i = 2) \\ & + Y_i(D_i = 0, T_i = (0,\ 0,\ 0)) \times P(D_i = 0|Z_i = 2), \end{aligned} \tag{5}$$

the distribution of the outcome under baseline for compliers multiplied by the population share of compliers and the distribution of the outcome under baseline for never-takers multiplied by the population share of never-takers. This implies that the outcome distribution for compliers under baseline is equal to

$$\begin{aligned} & Y_i(D_i = 0, T_i = (0, 1, 2)) = \\ & \qquad \frac{Y_i(Z_i = 0, D_i = 0) - Y_i(D_i = 0, T_i = (0, 0, 0)) \times P(D_i = 0|Z_i = 2)}{P(D_i = 2|Z_i = 2)}. \end{aligned} \tag{6}$$

We can then write the effect of the treatment for individuals who received the treatment instead of the baseline as

$$\tau_{\text{bt}} = \mathbb{E}[Y_i(D_i = 2, T_i = (0, 1, 2)) - Y_i(D_i = 0, T_i = (0, 1, 2))] \tag{7}$$

$$= \mathbb{E}\left[Y_i(D_i = 2, T_i = (0, 1, 2)) - \frac{Y_i(Z_i = 0, D_i = 0) - Y_i(D_i = 0, T_i = (0, 0, 0)) \times P(D_i = 0|Z_i = 2)}{P(D_i = 2|Z_i = 2)} \right] \tag{8}$$

$$= \mathbb{E}\left[Y_i(Z_i = 2, D_i = 2) - \frac{Y_i(Z_i = 0, D_i = 0) - Y_i(Z_i = 2, D_i = 0) \times P(D_i = 0|Z_i = 2)}{P(D_i = 2|Z_i = 2)} \right]. \tag{9}$$

The last equality is true because we can infer $Y_i(D_i = 2, T_i = (0, 1, 2))$ from the subpopulation with $(Z_i, D_i) = (2, 2)$ and $Y_i(D_i = 0, T_i = (0, 0, 0))$ from the subpopulation with $(Z_i, D_i) = (2, 0)$ since all these individuals are known to be compliers and never-takers, respectively. The sample analog of the expectation in equation (9) is

$$\widehat{\tau_{\text{bt}}} = \frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 2, D_i = 2\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 2\}} - \frac{\frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 0, D_i = 0\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 0, D_i = 0\}} - \frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 2, D_i = 0\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 0\}} \times \frac{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 0\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2\}}}{\frac{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 2\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2\}}}, \tag{10}$$

which, after some simplification, yields the local average treatment effect estimator of Angrist et al. (1996):

$$\widehat{\tau_{\text{bt}}} = \frac{\frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 2\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2\}} - \frac{\sum_{i=1}^{N} Y \cdot \mathbf{1}\{Z_i = 0\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 0\}}}{\frac{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2, D_i = 2\}}{\sum_{i=1}^{N} \mathbf{1}\{Z_i = 2\}}}. \tag{11}$$

The expression in the numerator represents the intent-to-treat effect, and the denominator represents the share of compliers in the sample.

3.3 *ML Estimator*

As discussed above, the estimators for τ_{pt} and τ_{bt} introduced in the two previous subsections asymptotically converge to the same quantity when the effect of the placebo on the outcome of interest is zero. Here, we exploit this fact to derive a more efficient ML estimator of the effect of the treatment on the treated for the case of binary outcomes in the context of a voter mobilization field experiment. The extension to continuous outcomes is in principle straightforward.

Our empirical analysis considers a voter mobilization campaign in which canvassers call or visit registered voters in an effort to encourage them to cast a ballot. The canvassers attempt to contact subjects assigned to the treatment or placebo but may not be able to do so. When contact is made, canvassers either deliver a message designed to encourage voting (the treatment) or recycling (the placebo).

The (aggregate) data generated by this field experiment are listed below. Note that these counts are sufficient statistics for the three-group design.

- n_{B} = # in baseline group,
- v_{B} = # voting in baseline group,
- n_{T}^{C} = # contacted in treatment group,
- v_{T}^{C} = # voting among those contacted in treatment group,
- $n_{\text{T}}^{\bar{C}}$ = # not contacted in treatment group,
- $v_{\text{T}}^{\bar{C}}$ = # voting among those not contacted in treatment group,
- n_{P}^{C} = # contacted in placebo group,
- v_{P}^{C} = # voting among those contacted in placebo group,
- $n_{\text{P}}^{\bar{C}}$ = # not contacted in placebo group,
- $v_{\text{P}}^{\bar{C}}$ = # voting among those not contacted in placebo group.

We now introduce some additional notation for the four parameters to be estimated (see Table 3). The first parameter is, α, the share of compliers in the population, which in our empirical application is the same as the probability of successfully contacting a voter. It is the same across all groups. The second parameter is the probability that a voter contacted in the treatment group, that is, a complier, votes $\left(\pi_{\text{T}}^{C}\right)$. The third parameter is the probability that a voter who could not be contacted, that is, a never-taker, votes $\left(\pi^{\bar{C}}\right)$. It is the same across all groups. The last parameter is the treatment effect on the treated (τ), that is, the difference in the probability of voting between a complier in the treatment group and a complier in either the baseline or the placebo groups. In addition to assumptions (a)–(f), we also assume that the placebo has no effect on turnout. We return to this assumption below.

Table 3 Notation

Parameter	*Definition*	*Description*
α	$P(D = 1\|Z = 1) \equiv P(D = 2\|Z = 2)$	Contact probability in placebo group ≡ contact probability in treatment group ≡ population share of compliers
π_{T}^{C}	$P(Y = 1\|Z = 2, D = 2)$	Probability of voting given treatment and contact
$\pi^{\bar{C}}$	$P(Y = 1\|Z = 1, D = 0) \equiv P(Y = 1\|Z = 2, D = 0)$	Probability of voting given no contact in placebo group ≡ probability of voting given no contact in treatment group
τ_{pt}	$P(Y = 1\|D = 2) - P(Y = 1\|D = 1)$	Treatment effect in placebo-treatment comparison
τ_{bt}	$P(Y = 1\|Z = 2, D = 2) - P(Y = 1\|Z = 0, D_z = Z)$	Treatment effect in baseline-treatment comparison
$\tau \equiv \tau_{\mathrm{pt}} \equiv \tau_{\mathrm{bt}}$		Treatment effect in three-group design

3.4 *Log-Likelihood Function*

Let $\log \mathcal{L}_{\mathrm{B}}$, $\log \mathcal{L}_{\mathrm{T}}$, and $\log \mathcal{L}_{\mathrm{P}}$ denote the log-likelihoods for the baseline, treatment, and placebo groups, respectively. The overall log-likelihood function $\log \mathcal{L}$ is given by

$$\begin{aligned}\log\mathcal{L} &= \log\,\mathcal{L}_{\mathrm{B}} + \log\,\mathcal{L}_{\mathrm{T}} + \log\,\mathcal{L}_{\mathrm{P}}\\ &= v_{\mathrm{B}}\log\left[\alpha\left(\pi_{\mathrm{T}}^{C}-\tau\right)+\left(1-\alpha\right)\pi^{\bar{C}}\right]+\left(n_{\mathrm{B}}-v_{\mathrm{B}}\right)\log\left[1-\left(\alpha\left(\pi_{\mathrm{T}}^{C}-\tau\right)+\left(1-\alpha\right)\pi^{\bar{C}}\right)\right]\\ &\quad + v_{\mathrm{T}}^{C}\log\left(\alpha\pi_{\mathrm{T}}^{C}\right)+\left(n_{\mathrm{T}}^{C}-v_{\mathrm{T}}^{C}\right)\log\left(\alpha\left(1-\pi_{\mathrm{T}}^{C}\right)\right)+v_{\mathrm{T}}^{\bar{C}}\log\left(\left(1-\alpha\right)\pi^{\bar{C}}\right)\\ &\quad + \left(n_{\mathrm{T}}^{\bar{C}}-v_{\mathrm{T}}^{\bar{C}}\right)\log\left(\left(1-\alpha\right)\left(1-\pi^{\bar{C}}\right)\right)+v_{\mathrm{P}}^{C}\log\left(\alpha\left(\pi_{\mathrm{T}}^{C}-\tau\right)\right)+\left(n_{\mathrm{P}}^{C}-v_{\mathrm{P}}^{C}\right)\log\left(\alpha\left(1-\left(\pi_{\mathrm{T}}^{C}-\tau\right)\right)\right)\\ &\quad + v_{\mathrm{P}}^{\bar{C}}\log\left(\left(1-\alpha\right)\pi^{\bar{C}}\right)+\left(n_{\mathrm{P}}^{\bar{C}}-v_{\mathrm{P}}^{\bar{C}}\right)\log\left(\left(1-\alpha\right)\left(1-\pi^{\bar{C}}\right)\right).\end{aligned} \tag{12}$$

ML estimates of the four parameters given the sufficient statistics are found by maximizing the function above. Of particular interest is the estimate of the treatment effect, $\hat{\tau}$. The estimated covariance matrix and hence standard errors of the parameter estimates follow from standard asymptotic theory (Cox and Hinkley 1974).

4 Empirical Application

A field experiment was conducted in August 2008. During the day preceding the Michigan primary elections voters who, according to Michigan's Qualified Voter File, had voted in November 2004 and 2006 but not in August 2006 were randomly assigned to one of three groups.[8] Members of the treatment group ($N = 8448$) were called by an automated dialing machine that conveyed the following prerecorded message: "We are calling to remind you to vote in tomorrow's election. Primary elections are important, but many people forget to vote in them. According to public records, you did vote in both November 2004 and 2006, but you missed the 2006 August primary. Please remember to vote tomorrow, August 5th.

[8]To sidestep statistical issues associated with clustered random assignment, we focus attention solely on one-voter households. Treatment effects are similar for two-voter households.

Press <1> if you would like us to provide future reminders like this one. Or press <2> if you would like your phone number removed from our list.'' This voter mobilization treatment is patterned after the ''social pressure'' mailings described by Gerber et al. (2008).

The placebo group ($N = 8357$) was called at the same time and from the same location. Following Nickerson (2008), the placebo script focused on recycling. The prerecorded message read as follows: ''We are calling to remind you of the importance of recycling. Michigan has the lowest recycling rate for plastics and other consumer products of any Great Lakes state, wasting energy and harming the environment. Please remember to separate your trash and recycle. Press <1> if you would like us to provide future reminders like this one. Or press <2> if you would like your phone number removed from our list.''

The remainder of the sample was allocated to the baseline group ($N = 304{,}948$). No calls were directed at this group.

The automated dialer recorded whether the call was received and how many seconds the listener spent on the line before hanging up. No messages were left on answering machines. We consider a successful contact to be a connection of any duration; to define contact as successful completion of the script risks violations of the perfect blindness assumption. By this conservative definition, 47.18% of the treatment group was contacted compared with 46.74% of the placebo group. Consistent with the perfect blindness assumption, this difference in contact rates is small and statistically insignificant ($\chi^2_1 = 0.314$, $p = .58$). We also compared key covariate distributions of individuals who received the placebo and individuals who received the treatment. We used χ^2_1 tests for differences in proportions for turnout in the 2002 general elections and the 2002 and 2004 primary elections ($p = .34$, .46, and .95). These results support the appropriateness of the perfect blindness assumption in our study.

Table 4 reports voter turnout rates for all three groups, measured using public records collected by registrars of voters. The turnout rates in the three experimental groups are 16.98% in the baseline group, 17.04% in the placebo group, and 17.85% in the treatment group. Turnout in the baseline group is indistinguishable from turnout in the placebo group ($p = .89$), which confirms that the placebo did not affect turnout. Note that among individuals assigned to the placebo group, turnout is noticeably higher for those successfully contacted (19.25%) than for those not contacted (15.10%). This demonstrates that compliance behavior is related to turnout. An observational study that simply compared the turnout of individuals contacted and not contacted by a campaign could reach quite misleading conclusions about the effectiveness of voter mobilization campaigns (see also Arceneaux et al. 2006).

Table 5 presents the ML estimates of the key parameters of interest, the most important of which is the treatment effect on the treated. The baseline-treatment comparison (column 2) generates an estimated treatment effect of 1.85 percentage points with a standard error of 0.89. The placebo-treatment comparison (column 3) yields an estimated treatment effect of 2.27 percentage points with a standard error of 0.91. The gain in efficiency associated with using the three-group design instead of the two-group designs is substantial. The three-group estimate (column 1) is 2.18 with a standard error of 0.75. In other words, augmenting the placebo-treatment design with a baseline group lowers the standard error from 0.91 to 0.75, an 18% reduction. To achieve a similar improvement in precision within the framework of a placebo-treatment two-group design would have required a 47% increase in sample size.[9]

[9]The ML estimator of τ can be understood as a weighted average of the two two-group estimators. We can recover the weight, w, by solving the following equation for w: $\hat{\tau} = w \cdot \hat{\tau}_{\text{pt}} + (1-w) \cdot \hat{\tau}_{\text{bt}}$. In our application, the ML estimator gives more weight to the treatment effect estimate derived from the placebo-treatment comparison than the estimate derived from the baseline-treatment comparison ($w = .786$).

Table 4 Experimental outcomes for voter mobilization study

	Baseline	*Placebo*	*Treatment*
Whole sample			
N	304,948	8357	8448
Contacted	—	3906	3986
Contacted (%)	—	46.74	47.18
Voted	51,766	1424	1508
Voted (%)	16.98	17.04	17.85
Contacted			
Voted	—	752	858
Voted (%)	—	19.25	21.53
Not contacted			
Voted	51,766	672	650
Voted (%)	16.98	15.10	14.57

Note. The table shows turnout in the voter mobilization study for the baseline, placebo, and treatment groups. The first row shows sample sizes. The second and third rows show the number and proportion of individuals contacted in each of the three experimental groups. Rows 4 and 5 show the number and proportion of individuals voting in the 2008 Michigan primary elections. The second part of the table conditions on successful contact; it shows the number and proportion of individuals voting in the 2008 Michigan primary elections within the subgroup of contacted individuals. The third part of the table conditions on no contact; it shows the number and proportion of individuals voting in the 2008 Michigan primary elections within the subgroup of individuals who were not contacted.

A χ^2 test can be used to assess the goodness of fit of the model to the data. Here, $\chi^2_3 = 0.86$, $p = .84$, indicating an extremely good fit.[10] Lack of fit would suggest a violation of at least one of the identifying assumptions of the three-group estimator (such as perfect blindness, the exclusion restriction for the instrument, or the absence of placebo effects).

Substantively, the results presented here contrast sharply with those obtained by prior experimental studies of prerecorded voter mobilization phone calls. Several large-scale prior experiments found prerecorded voter mobilization calls to have negligible effects on voter turnout, and none found significant positive effects (Green and Gerber 2008). No prior study, however, tested the effectiveness of calls that apply what Gerber et al. (2008) term ''social pressure.'' The phone calls tested here disclosed that voting is a matter of public record, and the results suggest that this message may be unusually effective.

5 Efficiency of the Three-Group Design

5.1 *Simulation Results*

We conducted a small simulation study that illuminates the conditions under which a three-group design is particularly advantageous. Replacing the data from our sample in which the proportion of compliers (the contact rate) was around 47%, we created new data sets by

[10]We have the following constraints: in the baseline group, there are two cells that must add to the size of the baseline group; in the treatment and placebo groups, there are four cells each that must add to the total sizes (contacted and vote, contacted and no vote, not contacted and vote, and not contacted and no vote). So, we have a total of $1 + 3 + 3 = 7$ degrees of freedom. Since we estimate four parameters, we end up with $7 - 4 = 3$ degrees of freedom for the χ^2 test.

Table 5 ML estimates with standard errors in parentheses

	Estimates using baseline, placebo, and treatment groups	*Estimates using baseline and treatment groups*	*Estimates using placebo and treatment groups*
$\hat{\alpha}$ (contact rate)	0.4697 (0.0038)	0.4718 (0.0054)	0.4696 (0.0038)
$\hat{\pi}_T^C$ (probability of voting given treatment and contact)	0.2153 (0.0065)	0.2153 (0.0065)	0.2153 (0.0065)
$\hat{\pi}^C$ (probability of voting given no contact)	0.1487 (0.0032)	0.1457 (0.0053)	0.1483 (0.0038)
$\hat{\tau}$ (treatment effect)	0.0218 (0.0075)	0.0185 (0.0089)	0.0227 (0.0091)

varying the proportion of compliers from .1 to .9, holding the other three model parameters equal to their point estimates shown in Table 5, column 1, and setting all ''observed'' counts equal to their expected values. In other words, the sample sizes for all three groups, the turnout rates for compliers and never-takers, and the treatment effect were all held constant, but the admixture of compliers and never-takers was systematically varied.

Figure 1 compares the standard errors for the estimated treatment effects from the baseline-treatment, placebo-treatment, and three-group estimators as a function of the contact rate (α). The first thing to note is that standard errors from the three-group design are always smaller than standard errors from the two-group designs. In other words,

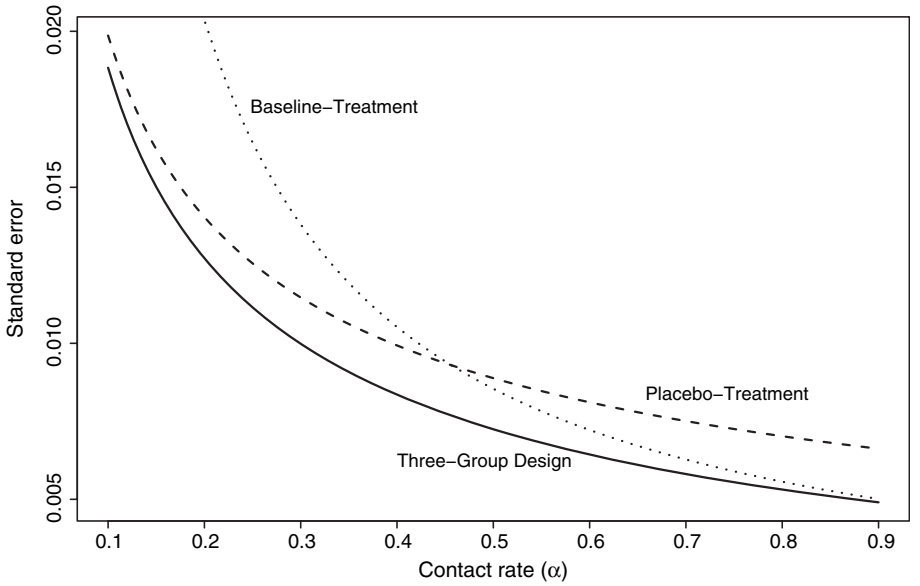

Fig. 1 Estimated standard errors as a function of the contact rate. The graph shows standard errors for the estimated treatment effect from the baseline-treatment, placebo-treatment, and three-group designs as the contact rate (α) ranges from .1 to .9. All other parameters are identical to the estimates presented in Table 5, column 1.

the three-group design is strictly more efficient than either of the two-group designs. Although this in itself is not surprising given that the three-group design enjoys a larger overall sample size than either two-group design, its advantage over the other designs varies with the contact rate. For very low contact rates, standard errors from the three-group design are not much smaller than standard errors from the placebo-treatment design. For very high contact rates, standard errors from the three-group design are not much smaller than standard errors from the baseline-treatment design.

Figure 2 rescales the results presented in Fig. 1 to depict the relative sizes of the standard errors from the three designs more directly. The dashed line shows the ratio of the standard errors for the estimated treatment effects from the placebo-treatment and three-group designs. The dotted line shows the ratio of the standard errors for the estimated treatment effects from the baseline-treatment and three-group designs. The contact rate (α) still ranges from .1 to .9. We can see that for very low contact rates, the ratio of the standard errors from the placebo-treatment and three-group designs approaches 1, that is, standard errors from the placebo-treatment design are only slightly larger than standard errors from the three-group design. The ratio of the standard errors from the baseline-treatment and three-group designs, on the other hand, increases dramatically, illustrating the advantage of the three-group design over the baseline-treatment design when contact rates are low. When contact rates are very high, we find the opposite result. Standard errors from the baseline-treatment design are only slightly larger than standard errors from the three-group design (their ratio approaches 1), whereas the advantage of the three-group design over the placebo-treatment design becomes more pronounced. The height of the gray area below the lower of the two curves denotes the improvement in efficiency associated with the use of the three-group design instead of the second-best design, that is, the baseline-treatment design for low contact rates and the placebo-treatment design for high contact rates.

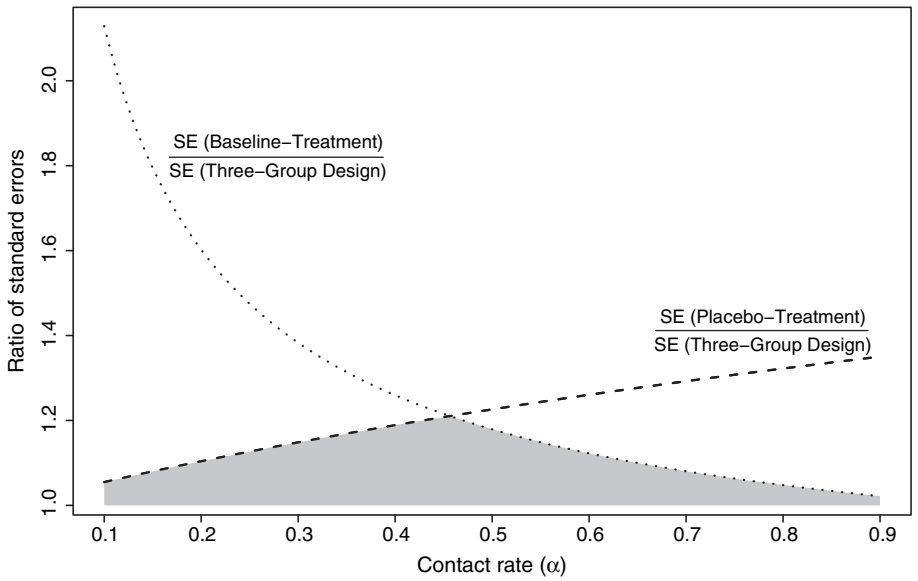

Fig. 2 Ratio of estimated standard errors as a function of the contact rate. The graph shows the ratios of the standard errors for the estimated treatment effects from the baseline-treatment and three-group design and the placebo-treatment and three-group design as the contact rate (α) ranges from .1 to .9. All other parameters are identical to the estimates presented in Table 5, column 1.

5.2 *Optimal Design*

To this point, we have considered the case in which a placebo-treatment design is augmented by the inclusion of a baseline group; our focus has been the efficiency gains of a three-group design. Here, we consider a different question: Suppose the baseline group were obtained at no cost, but the researcher must pay the same cost for each additional treatment or placebo observation. In our empirical application, for example, the baseline group was obtained at no cost, but we paid a fixed rate for each completed phone call, regardless of the script. In this case, how should a researcher allocate resources to maximize the precision with which the treatment effect is estimated?

Again using the estimates from Table 5, column 1 as the basis for our simulations, Fig. 3 shows standard errors from the three-group design as a function of the contact rate and the proportion of subjects assigned to the treatment group (out of the fixed pool of subjects assigned to either placebo or treatment). The proportion of subjects assigned to the treatment group ranges from .4 to 1; five curves show standard errors for five different values for the contact rate. The minimum standard error for each value of the contact rate is denoted by an empty circle. We can see that as the contact rate increases from .1 to .5, the proportion of subjects assigned to the treatment group that minimizes the standard error increases. With a contact rate of .5 or higher, moving the whole placebo group into the treatment group (which means simply using a baseline-treatment design) becomes the most efficient design choice, at least given the parameter values on which this particular set of simulations is based. Note that *ex post*, given the estimated proportion of compliers $\left(\hat{\alpha} = 0.4697\right)$ and the other estimated parameters in our empirical example, it would have been slightly

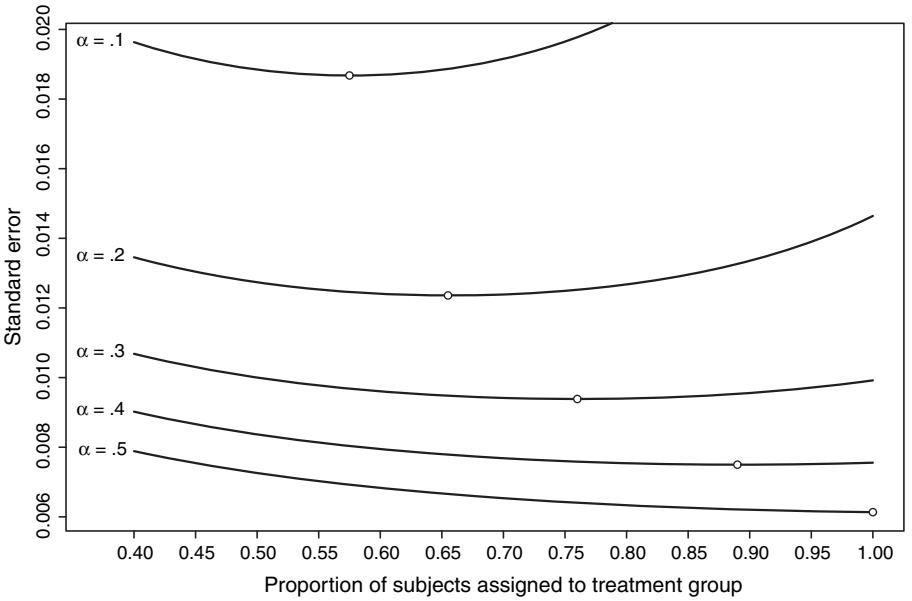

Fig. 3 Estimated standard errors as a function of the contact rate and the proportion of subjects assigned to the treatment group. The graph shows standard errors for the estimated treatment effect from the three-group design for five different values for the contact rate (α) as the proportion of subjects assigned to the treatment group (out of the total number of subjects assigned to treatment and placebo) ranges from .4 to 1. All other parameters are identical to the estimates presented in Table 5, column 1. Empty circles denote minima.

more efficient to move the whole placebo group into the treatment group and to use a baseline-treatment design instead. *Ex ante*, however, with the proportion of compliers, the treatment effect, and the turnout rate among compliers and never-takers unknown, the three-group design provides an insurance against situations in which two-group designs perform extremely poorly.[11]

6 Discussion

Ready availability of outcome data for baseline groups offers researchers a low-cost opportunity to improve the efficiency of their experimental designs. Costless access to outcome data for the baseline group increased our experiment's effective sample size by almost half. Studies with a larger population share of compliers can expect to make even more valuable use of the baseline group by assigning more individuals to the treatment group than the placebo group.

Another virtue of the three-group design is that it helps detect and correct unexpected problems. The three-group design enables researchers to verify that the placebo has no effect on the outcome of interest (Rosenthal 1985). If the treatment were accidentally administered to the placebo group or if the placebo somehow affected the outcome of interest, the placebo-treatment comparison would produce biased results. In the absence of a baseline group, this bias would be difficult to detect empirically. Indeed, in medical research tampering with placebo groups has sometimes become evident only decades after placebo-treatment comparisons yielded misleading results (Silverman 1980). The three-group design enables researchers to compare the placebo and baseline groups to verify that the placebo was ineffectual. Moreover, the three-group design is superior to a baseline-treatment design when the population share of compliers is low. Given that researchers are often uncertain about placebo effects, perfect blindness, and the extent of noncompliance with treatment assignment, the three-group design is akin to an insurance policy. It may not necessarily prove to be an optimal allocation of resources *ex post*, but *ex ante* it guards against significant risks.

The class of experiments to which the three-group design might be applied is potentially quite broad. Especially relevant are so-called ''crossover designs'' (Shadish et al. 2002, chap. 8) in which a random sample of n observations is drawn from a population of size N; m observations are assigned to a group that receives a treatment immediately, whereas $n - m$ observations are assigned to a group that receives the same treatment at some later date. Outcomes are measured for all N observations prior to the administration of the treatment to the second group; this group functions as a placebo group. A recent prominent example of such a crossover design is Progresa, the Mexican program for education, health, and nutrition (Gertler 2004).

Although crossover experiments may be attractive to participants who would otherwise balk at participating in a study that may place them in a group that never receives the treatment, crossover designs may nonetheless confront problems of noncompliance. For example, n randomly selected observations might be encouraged to participate in a program with the stipulation that the researcher will determine the period during which the treatment is administered. Many subjects may be unavailable, uninterested, or unwilling to

[11] Additional simulation results for other combinations of parameter values are available in an online appendix. These results demonstrate that even with a large population share of compliers, it is not always optimal to switch to a baseline-treatment design, depending on the exact combination of parameter values.

participate under these conditions. The advantage of the three-group design is that it enhances the efficiency of a placebo-treatment comparison and provides a means for detecting placebo effects, which could arise in the context of crossover designs if subjects change their behavior in anticipation of future treatment. Possible applications include crossover experiments whose outcomes are readily measured for all N observations. In the social sciences, such outcomes may be obtained from administrative records for individuals (births, deaths, bankruptcies, convictions, and campaign contributions), organizations subject to public disclosure requirements (tax statements, board composition, and funding sources), and political entities (budgets, policies, portfolios, and decisions). Three-group designs are likely to become much more common as these research opportunities are explored.

Funding

The Institution for Social and Policy Studies at Yale University; Daniel Rose Fund supporting the Technion-Yale Initiative in Homeland Security and Counterterror Operations Research to EHK.

References

Abadie, Alberto. 2003. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* 113(2):231–63.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–55.

Angrist, Joshua D., and Jörg-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis* 14(1):37–62.

Boruch, Robert F. 1997. *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: SAGE Publications.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90(430):443–50.

Cheng, Jing, and Dylan S. Small. 2006. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society, Series B* 68(5):815–36.

Cox, David R., and David V. Hinkley. 1974. *Theoretical statistics*. London: Chapman and Hall.

de Craen, Anton J. M., Ted J. Kaptchuk, Jan G. P. Tijssen, and J. Kleijnen. 1999. Placebos and placebo effects in medicine: Historical overview. *Journal of the Royal Society of Medicine* 92(10):511–5.

Frangakis, Constantine E., and Donald B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58(1):21–9.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. The illusion of learning from observational research. In *Problems and methods in the study of politics*, eds. Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud, 251–73. Cambridge: Cambridge University Press.

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

Gertler, Paul. 2004. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review* 94(2):336–41.

Green, Donald P., and Alan S. Gerber. 2008. *Get out the vote: How to increase voter turnout*. 2nd ed. Washington, DC: Brookings Institution Press.

Efron, B., and D. Feldman. 1991. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* 86(413):9–17.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–60.

Imbens, Guido W. 2007. Nonadditive models with endogenous regressors. In *Advances in economics and econometrics*. Vol. III. Chapter 2, eds. Richard Blundell, Whitney Newey, and Torsten Persson, 17–46. Cambridge: Cambridge University Press

Imbens, Guido W., and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–76.

Imbens, Guido W., and Paul R. Rosenbaum. 2005. Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society, Series A* 168(1):109–26.

Imbens, Guido W., and Donald B. Rubin. 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 25(1):305–27.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference*. Cambridge: Cambridge University Press.

Nickerson, David W. 2005. Scalable protocols offer efficient design for field experiments. *Political Analysis* 13(3):233–52.

———. 2008. Is voting contagious? Evidence from two field experiments. *American Political Science Review* 102(1):49–57.

Rosenthal, Robert. 1985. Designing, analyzing, interpreting, and summarizing placebo studies. In *Placebo: Theory, research, and mechanisms*, eds. Leonard White, Bernard Tursky, and Gary E. Schwartz, 110–36. New York: Guilford Press.

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

———. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2(1):1–26.

———. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6(1):1–26.

———. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 5(4):472–80.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Silverman, William A. 1980. *Retrolental fibroplasia: A modern parable*. New York: Grune.

Torgerson, David J., and Carole J. Torgerson. 2008. *Designing randomized trials in health, education and the social sciences*. New York: Palgrave Macmillan.