Advance Access publication May 11, 2005

Political Analysis (2005) 13:233–252
doi:10.1093/pan/mpi015

Scalable Protocols Offer Efficient Design for Field Experiments

David W. Nickerson
Department of Political Science, University of Notre Dame,
217 O'Shaughnessy Hall, Notre Dame, IN 46556
e-mail: dnickers@nd.edu

Experiments conducted in the field allay concerns over external validity but are subject to the pitfalls of fieldwork. This article proves that scalable protocols conserve statistical efficiency in the face of problems implementing the treatment regime. Three designs are considered: randomly ordering the application of the treatment; matching subjects into groups prior to assignment; and placebo-controlled experiments. Three examples taken from voter mobilization field experiments demonstrate the utility of the design principles discussed.

1 Introduction

The past few years have witnessed a growth in the popularity of field experiments as a research methodology for studying political behavior. The ability to establish clear causal relationships is the primary attraction of randomized experiments. Experiments obviate the need for complicated modeling assumptions through the controlled manipulation of a variable of interest. Unfortunately, this control often removes analysis from the real world and places it within the artificial confines of the laboratory, creating questions about the external validity of the findings. Field experiments can ameliorate some concerns about external validity, but control over the execution of the protocol is diminished and unexpected problems can arise. Solutions to problems encountered in the field are typically labor intensive and expensive. This article argues that scalable experimental designs preserving statistical efficiency can free up resources needed to address problems in execution.

An experiment, at its most basic, randomly divides subjects into two groups: one that receives the factor of interest (the treatment group) and one that is not exposed to the factor (the control group). Since assignment to each group is random, any systematic difference between the two groups in the dependent variable is, on average, attributable to the variable of interest. Problems arise when there is a failure to treat the members of the treatment group or the treatment is inadvertently applied to the control group. Failure to treat can be manifested in many different ways but falls into three broad categories: delivery, the researcher's ability to distribute the treatment accurately; receipt, the subject's ability to receive the treatment; and adherence, whether or not the subject

Author's note: The author would like to thank Donald Green, Alan Gerber, and three anonymous reviewers for their helpful comments.

obeys the prescribed treatment regime (Lichstein et al. 1994). While failure to treat decreases statistical efficiency, it does not inherently bias estimates. However, concerns about protocol execution often result in subject attrition, which can bias experimental estimates.

Adjusting for such problems usually takes place after the fact through statistical modeling, but the process involves imposing a series of assumptions that may not accurately capture the data generating process. Since the validity of these modeling assumptions is often untestable, the value of the experiment over research based upon observational research is questionable when compliance with the experimental protocol and subject attrition must be modeled (see Heckman and Smith 1995).

An alternative strategy is to design experimental protocols that preserve precision by implementing scalable experiments. Most problems in the field can be addressed with additional time and money, which are in limited supply for most studies. This article presents three experimental designs that can be scaled in response to problems encountered in the field and demonstrates the improvement in statistical efficiency through a series of propositions and real-world examples. The contribution of this article lies not in the designs themselves (which are all commonly used throughout experimental sciences[1]), but in the discussion and examples from political behavior of how shifting resources in response to field contingencies can maximize statistical efficiency.[2]

The statistical power behind an experiment is a function of the variance of its estimated treatment effect. The variance of the treatment effect is a function of three primary components subject to experimental control: the number of subjects in the experiment, the ratio of subjects assigned to the treatment and control groups, and the percentage of the subjects to whom the correct treatment was actually applied (application rate).[3] Problems in the field usually adversely affect the application rate (i.e., members of the treatment group are not given the treatment or the control group accidentally receives the treatment). This article explains how three different experimental protocols preserve statistical efficiency by shifting shortfalls in the application rate to other components of the standard error formula. First, protocols that randomly determine the order in which the treatment is applied to subjects allow the researcher to adjust the treatment/control ratio to account for resource shortfalls. Second, protocols that match subjects within groups prior to random assignment into the treatment and control conditions place the burden of shortfalls squarely on the number of subjects in the experiment rather than the application rate. Brief algebraic proofs are provided to demonstrate the superiority of the two types of protocols to experimental designs that lower the application rate of the treatment. Finally, the relative statistical efficiency of placebo-controlled experiments will also be considered and discussed. Under some conditions, which are outlined below, placebo-controlled experiments are more efficient than protocols that randomize the order of treatment or match subjects.

The examples provided in this article will be taken from voter mobilization field experiments. Voter mobilization is unusual in that delivery of, receipt of, and adherence to the treatment regime occur simultaneously. The knock at the door encouraging a person to

[1]Four good reference books for experimental design are Wu and Hamada (2000), Montgomery (2001), Riffenburgh (1998), and Box et al. (1978).

[2]Leslie Kish (1965) discusses a nearly identical notion he calls "design effects." The context of Kish's discussion is survey research, but the logic applies equally well to the context of field experimentation.

[3]By collecting control variables, the researcher can lessen the unexplained variance of the estimand but cannot actually minimize the outcome variance itself. Control variables are an extremely useful tool and can be used in conjunction with the protocol designs discussed herein.

vote (i.e., delivery) lasts only a few seconds, so the subject has little opportunity not to hear the message (i.e., receipt) and nothing further to do (i.e., adherence).[4] As a result, voter mobilization experiments constitute a very clean set of illustrations of the design principles because application of the treatment is captured by a single concept and number—namely, the contact rate.[5] Other types of field experiments will likely be somewhat more complicated in this regard, but the results and design principles will still hold.

The article will begin by briefly highlighting a few pertinent facets of designing and analyzing field experiments. Second, the chief drawback to the standard experimental design will be described and explained. The discussion will then turn to a more efficient experimental protocol where the order in which subjects are treated is randomized (the rolling protocol). Unfortunately, there are few settings where it is possible to fully randomize the order in which subjects are treated, so the fourth section of the article describes a more general design where treatment and control subjects are paired together (the matched protocol). Because the major source of inefficiency is the application rate, a design that uses the application of a placebo as a comparison group for those treated is also considered. The article concludes by discussing the relative costs of each experimental design and complicating factors.

2 A Brief Introduction to Field Experiments

The utility of randomized controlled experiments was noted by R. A. Fisher in 1935 and took root in the natural and social sciences in the decades that followed. An experiment, in its most basic form, is any study in which one or more treatments are applied to subjects randomly (see Rubin 1974, p. 689, for a more lengthy discussion). The random assignment of the variable of interest assures (on average) that confounding factors, such as measurable and immeasurable causes of the dependent variable, are the same in the treatment and control groups.[6] This balance between treatment and control groups allows for very easy assessment of the effect of the treatment. To calculate the intent-to-treat effect (ITT), simply subtract the observed average value of the dependent variable of the control group from the observed rate in the treatment group.

This logic can easily be expressed mathematically. Let T equal the assignment to the treatment condition (i.e., $T = 0$ implies assignment to the control group and $T = 1$ implies assignment to the first treatment group; $Y_i^{T=1}$ and $Y_i^{T=0}$ represent the outcome measures for the individuals assigned to the treatment and control groups, respectively). Equation (1) represents the model for the intent-to-treat effect where B is the baseline level of the outcome variable (i.e., the sum of the observed and unobserved causes of Y) and δ is the effect of the treatment assignment upon the measured outcome,

$$Y_i = B_i + \delta T_i. \tag{1}$$

[4]Since voter mobilization experiments are often associated with existing political organizations, consent of the subjects is not a concern. The campaign would contact a set of residents regardless; the researcher simply randomizes existing activities.

[5]In the United States, voter turnout is recorded and publicly available. As a result, subject attrition due to measurement of the dependent variable (or lack thereof) is not a concern for voter mobilization experiments.

[6]It should be noted that endogeneity is another confounding factor that randomized experiments sidestep.

Since the assignment to treatment is random, $\lim_{i\to\infty} B_i^{T=1} - B_i^{T=0} = 0$ and the intent-to-treat effect can be calculated as follows:

$$\delta_{ITT} = \bar{Y}^{T=1} - \bar{Y}^{T=0}. \tag{2}$$

This estimated intent-to-treat effect provides an unbiased estimate of the effect of the assignment to a treatment regime.

While field experiments are common in public policy research and enjoyed brief popularity during the early 1950s, experiments have largely been confined to the laboratory regarding political behavior. The reason for the dearth of experimental studies of political behavior in the field is the difficulty of implementing protocols in the real world. One complication that arises is applying the correct treatment to the correct subject. Lichstein et al. (1994) divide treatment application into three separate components: treatment delivery, treatment receipt, and treatment adherence. In the field, each and every one of these parts of treatment can fail. If significant numbers of subjects in the treatment group fail to complete the assigned treatment regime, then the intent-to-treat analysis will understate the true effect of the treatment upon those subjects who actually received it. What is needed is an adjustment of the intent-to-treat effect to estimate the treatment effect upon those actually treated.

Happily, this transformation is easy to compute (see Angrist et al. 1996a and Gerber and Green 2000 for a full discussion). Failure to treat creates two types of experimental subjects: those who would have completed the assigned treatment regime and those who would not, or, compliers and noncompliers. Let B_C and $B_{\sim C}$ represent the baseline outcome measures for compliers and noncompliers, respectively. Let α represent the application rate of the experiment (i.e., percentage of individuals assigned to the treatment condition successfully receiving the treatment minus the percentage of the individuals assigned to the control group that inadvertently received the treatment). Since the experimental conditions are randomly assigned, both the treatment and control groups will contain an equal proportion of compliers and noncompliers on average (i.e., $E(\alpha^{T=1} - \alpha^{T=0}) = 0$).[7] It is now possible to represent the causal model as follows:

$$Y = \alpha(B_C + \delta T) + (1 - \alpha)B_{\sim C}. \tag{3}$$

Due to the random assignment, $\lim_{n\to\infty} \beta_C^{T=1} - \beta_C^{T=0} = 0$ and $\lim_{n\to\infty} \beta_{\sim C}^{T=1} - \beta_{\sim C}^{T=0} = 0$. Equation (4) solves for δ and provides an unbiased estimator for the treatment effect upon those subjects treated (TOT).

$$\delta_{TOT} = \frac{\bar{Y}^{T=1} - \bar{Y}^{T=0}}{\hat{\alpha}}. \tag{4}$$

The logic behind this straightforward estimator is that since the treatment and control groups are randomly assigned, the two will generally possess an equal proportion of persons who would comply with the treatment regime given the opportunity. That is, if 40% of the persons assigned to the treatment group complied with the protocol, a hypothetical application of the treatment regime to the control group would also yield

[7] One can use the sample to estimate α by calculating the percentage of subjects in the treatment group actually treated.

40% compliance on average.[8] It is important to note that the treatment effect being estimated is only for those subjects who comply with the prescribed protocol. There is no way of estimating the treatment effect for those subjects who do not or would not participate without imposing stronger restrictions (Angrist et al. 1996b). Narrowly construing the object of estimation to be the average treatment effect upon those treated makes failure to treat a problem of external validity (i.e., how the treatment affects those subjects in the treatment group not treated) rather than one of bias.[9]

The concern of this article is the precision of the estimated treatment effect upon those treated. Equation (5) presents the formula for the variance of the estimated treatment effect upon the treated subjects from Eq. (4):

$$Var(\delta_{TOT})_E = \frac{\sigma^2}{A^2 NT(1-T)}, \tag{5}$$

where σ^2 = the variance of the estimand, Y; A = the application rate (i.e., the percentage of subjects in the treatment group to whom the treatment is actually applied minus the percentage of subjects in the control group to whom the treatment is applied); N = the total number of subjects in the experiment; and T = the percentage of subjects assigned to the treatment group (with $1-T$ being assigned to the control group),[10] Minimizing the variance of the estimated treatment effect allows a researcher to speak more precisely about and be more confident in the results of an experiment. Equation (5) offers a few obvious conclusions. First, as the total number of subjects, N, in the experiment decreases, the variance increases. So, all else being equal, big experiments have more statistical power than small experiments. Second, since $0 \leq A \leq 1$, as the application rate decreases, the variance of the estimate increases. Thus, treated members of the control group and untreated members of the treatment group decrease the power of the experiment. Third, the further the proportion of subjects in the treatment group moves away from 0.5 (a 50–50 split between treatment and control), the greater the variance of the estimated treatment effect. An experimenter can control A, N, and T, so it is useful to establish the relative importance of each of these factors in the research design through two propositions.[11]

Proposition 1. *There is a trade-off between increasing the number of subjects in an experiment and lowering the application rate. That is, adding subjects to an experiment increases the statistical precision only when a sufficient proportion of the new subjects*

[8]The Wald estimator discussed in Angrist et al. (1996a) makes the following five assumptions: 1) the assignment to treatment and control conditions was random; 2) neither the assignment of nor compliance with the treatment regime for one subject changes the outcome or compliance for another subject; 3) assignment to the treatment condition increases the likelihood of receipt of the treatment; 4) the random assignment has no effect upon the experimental outcome, except through the application of the treatment; and 5) there are no subjects who would always reject the treatment if assigned but take the treatment when not assigned. These five assumptions are unlikely to be problematic for many topics in political behavior. However, there may be instances in which one of the assumptions is problematic and a researcher will need to revert to intent-to-treat analysis (see Shadish et al. 2002, p. 322).

[9]Donald Campbell has clearly and eloquently differentiated concerns of external validity (i.e., how do results hold over time, place, persons, etc.) from internal validity (i.e., whether the inferred relationship is causal) and bias (i.e., systematic error in the estimate or inference). Detailed discussions can be found in Campbell and Stanley (1963), Cook and Campbell (1979), and Shadish et al. (2002).

[10]This terminology will be used throughout this article.

[11]The benefit of adding subjects to an experiment (increasing N) always outweighs the balance between treatment and control ratio. This is because it adding subjects to either the treatment or the control group increases the precision of the experimental estimate.

can be given the correct treatment.[12] *Specifically, an experiment with N+n subjects will have more power than an experiment with N subjects only when the application rate does not decline by more than* $\sqrt{N/(N+n)}$. *(Proof in Appendix)*

Proposition 2. *There is a trade-off between the treatment/control ratio and the application rate. That is, subjects should be shifted from the control group to the treatment group only when the application rate is not significantly lowered. Specifically, moving n subjects from the control to the treatment group increases statistical power only when the application rate declines by less than* $\sqrt{t(N-t)/(t+n)(N-t-n)}$, *where t is the number of subjects in the treatment group. (Proof in appendix)*

Taken together, Propositions 1 and 2 hint at the importance of the application rate. Adding subjects to an experiment is good in general, but if the ability to apply the correct treatment is a function of the number of subjects (or the treatment/control ratio), then the benefit of the increased size may be offset by reduced contact rates.

This point has practical importance when conducting a field experiment. Managerial oversight is necessary for many field experiments, and the size of the experiment may exceed a researcher's ability to control the quality of protocol implementation. Similarly, applying an experimental treatment may be labor intensive, and increasing the number of subjects to receive the treatment may spread resources too thinly. For example, suppose a researcher wants to study the effects of public service announcements (PSAs) on teenage smoking. She sends a randomly selected set of radio stations the PSA. The treatment is more likely to be applied (i.e., the PSA aired) if the researcher develops relationships with station managers and DJs. While mailing the PSA to 50 stations may be just as easy as sending it to 10 stations, if the researcher only has the time to make introductory and follow-up calls to 10 stations (thereby ensuring that the PSAs are aired), she might be well advised to keep the treatment group at 10.[13] The application rate may not be a concern in a laboratory setting where the researcher has control over what occurs, but it demands considerable attention when conducting experiments in the field.

The importance of the application rate is more apparent when the vagaries of conducting field experiments are considered. Despite careful planning and meticulous attention to detail, experimental protocols can be difficult to administer in practice. Subject retention may be more expensive than anticipated, volunteers can fail to show up, bad weather can cancel events, insufficient materials may be available—the potential list of problems is endless. To quantify the extent of havoc problems can wreak upon field experiments, let $0 \leq P \leq 1$ represent the percent of the protocol executed faithfully, where 1 means everything went exactly as planned and 0 means there was no experiment to speak of. To focus this very abstract operationalization it may be useful to think of P as the percentage of planned workers who actually assist in conducting the experiment. For instance, if a researcher determines that she needs 20 assistants to apply the treatment but only 15 show up, then $P = 0.75$.

Virtually any problem with implementing an experimental protocol can be captured by the parameter P, because an increase in resources can solve most problems and resources are fungible. For instance, suppose a researcher asked subjects to read a newspaper on a daily basis and answer attitudinal questions periodically. Suppose further that the monetary incentive for compliance proved twice as small as necessary for most subjects to

[12] Assuming that the variance of the estimand for the subjects added is the same as the variance of the estimand for the original subjects.

[13] This depends upon how many radio stations would play the PSA with no call (i.e., the baseline application rate).

complete the program. The researcher, therefore, would be able to complete only half of the study (whether by cutting the treatment regime in half or recruiting half as many subjects) and $P = 0.5$. Almost every conceivable problem can be measured in this manner.

A quick discussion of bias is in order here. Noncompliance with the treatment regime does not necessarily introduce bias. On average, the control group will contain a similar proportion of compliers and noncompliers, so the estimated effect of the treatment upon the treated will be biased only when the assumptions of the Angrist et al. (1996a) approach are violated (see note 8). However, noncompliance with the treatment regime is often associated with difficulty in measuring the dependent variable, in which case the subject cannot be included in the analysis in any straightforward manner.[14] If attrition from the experiment is correlated with treatment (negatively or positively), then estimates may be biased.[15] Every effort should be made to ensure that subjects complete the course of the experiment.[16] Unfortunately, such efforts usually are time consuming and expensive (see Shadish et al. 2002, pp. 323–39, for a discussion of possible solutions) and may prove difficult to budget for. While scalable experimental protocols may not be able to solve the problem of bias from differential attrition, they can free the resources necessary to tackle such problems and preserve sufficient statistical efficiency to obtain meaningful results.

In the following sections it will be demonstrated that the statistical power of the standard experimental protocol suffers considerably from even small drops in P but that scaled protocols offer an efficient solution to problems encountered in the field.

3 Standard Experimental Protocol

In the simplest experimental protocol, the researcher randomly assigns the universe of subjects to treatment and control conditions and then applies the correct treatment to each group. The number of subjects to be examined, N, is set, as is the treatment/control ratio, T. The component of the experiment that may be susceptible to problems is the application rate, A. As stated earlier, field experiments will rarely have an application rate of 1—contacting and organizing citizens in their natural habitats is more difficult than treating psychology students in a laboratory. Problems in the field will lower the application rate below its baseline. Thus, the application rate for the standard experimental protocol is PA. For instance, a researcher conducting a voter mobilization experiment may anticipate contacting people at home 30% of the time ($A = 0.3$). However, if only half of the needed volunteers arrive to knock on doors ($P = 0.5$), only half of the doors in the treatment group will be attempted, and the doors that are attempted will still have a success rate of 30%. The effective application rate for the entire treatment group in the experiment becomes an abysmal 0.15. The influence of problems can be incorporated into the formula for the variance of the treatment effect upon the treated for the standard protocol.

$$Var(\delta_{TOT})_S = \frac{\sigma^2}{P^2A^2NT(1-T)} \tag{6}$$

As the percent of the protocol implemented decreases, P, the precision of the estimate decreases. In the standard experimental protocol, any decline in planned resources is magnified because the shortfall affects the application rate, which is a squared term.

[14]Techniques to cope with missing data do exist (see Little and Rubin 1987 or Allison 2002). However, all missing data techniques are not costless and impose additional assumptions on the analysis.
[15]Goodman and Blum (1996) conclude that few studies analyze attrition from experiments.
[16]Groves (1989) discusses the trade-offs between attrition and sample size.

A tempting but incorrect method of boosting the application rate is to shift untreated members of the original treatment group into the original control group (for examples in voter mobilization see Eldersveld 1956; Adams and Smith 1980; Miller et al. 1981). The practice leads to biased inferences, because nontreatment is likely to be correlated with the dependent variable.[17] However, the next section will discuss a protocol designed to allow precisely this shift of subjects without biasing the estimates.

4 Rolling Experimental Protocol[18]

Rather than divide the total group of subjects into firm treatment and control groups, subjects are placed into the random order in which they will be treated. Since the only difference between the first subjects and the last subjects to be treated is the random number, those subjects for whom the application of the treatment was never attempted can simply be shifted to the control group. It is very important to note that the subjects for whom application of the treatment was attempted should remain in the treatment group and not be moved to the control group. There are likely to be systematic differences between those subjects for whom the treatment cannot be applied and those for whom application is successful.

Under this rolling protocol, both the number of subjects in the experiment, N, and the rate of application, A, remain fixed. The only component of the experimental variance subject to problems in execution is the treatment control ratio, T. If everything goes as planned, a researcher places T subjects in the treatment group and $(1-T)$ subjects into the control group. However, if problems in execution arise and only a portion of the protocol is implemented, only PT subjects may have application of the treatment attempted and the rest are rolled into the control group. The variance of the estimated treatment effect for such a protocol can be expressed as

$$Var(\delta_{TOT})_R = \frac{\sigma^2}{A^2 NPT(1-PT)}. \tag{7}$$

Because P is strictly positive and less than or equal to one, Eq. (7) has the intuitive result that the overall variance of the estimator increases as P decreases. Comparing Eq. (6) to Eq. (7), it is possible to state the following proposition:

Proposition 3. *For $0 < P < 1$, the rolling protocol estimate of the treatment effect exhibits less variance (hence more statistical precision) than the estimate under the standard protocol. (Proof in appendix.)*

The intuition behind the proof is straightforward. Using the standard protocol, problems affect the application rate and are therefore squared when calculating the variance of the estimate. In contrast, the rolling protocol shifts any shortfalls onto the treatment/control ratio, which inflates the variance of the estimate less than the application

[17] It is important to note that the correlation may not be observed. Managing unobserved heterogeneity is the chief advantage of controlled experiments, and moving subjects based upon contact defeats this purpose.

[18] The rolling protocol is occasionally referred to as a "fully randomized" protocol since the timing of the treatment application is also randomly determined. However, reference books seldom refer to the procedure as its own experimental design. Instead, randomizing the sequence of treatment is presented as a general design principle (e.g., Wu and Hamada [2000, pp. 9–11]; Montgomery [2001, pp. 60–63]; and Box et al. [1978, p. 405]).

Table 1 Experimental Voter Mobilization Results under Different Protocol Designs

City	*Boston*		*Bridgeport*		*Denver*		
Column	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Protocol	*Rolling*	*Standard*	*Matched*	*Standard*	*Placebo*	*Matched*	*Standard*
Number of subjects in study	7055	7055	1900	19184	562	3398	11080
Percent assigned to treatment	17.1%	50.7%	50.0%	50.0%	50.4%	50.1%	50.0%
Application rate	55.4%	18.7%	27.9%	2.7%	100.0%	16.7%	5.1%
Control group voting rate	54.5%	54.4%	9.7%	12.2%	39.1%	38.4%	39.4%
Treatment group voting rate	56.1%	55.1%	13.6%	12.3%	47.7%	40.6%	40.3%
Intent to treat effect	1.6%	0.7%	3.9%	0.1%	8.6%	2.2%	0.9%
Estimated effect on the treated	2.9%	3.5%	13.9%	4.9%	8.6%	13.2%	17.7%
Standard error of estimate	2.8%	6.3%	4.9%	17.3%	4.1%	10.0%	18.2%

rate. When it is possible to randomize the order in which subjects receive the treatment, the rolling protocol is the optimal design for conserving statistical precision in the face of shortfalls.

A voter mobilization experiment conducted in Boston during the 2001 mayoral election offers an example of the superiority of the rolling protocol over the standard protocol.[19] Phone numbers were obtained for 7055 registered voters, and half of these people were randomly selected to be called the weekend before the election by volunteers. The phone numbers were then placed in a random order and volunteers called from the top of the list to the bottom. Callers reached the intended person 55% of the time at the numbers dialed. Unfortunately, fewer volunteers than expected showed up, and only 1209 of the 3577 numbers in the treatment group were attempted. Because the phone list was randomly ordered, those numbers not attempted could be shifted into the control group, thereby preserving the 55% rate of application, but the treatment/control ratio fell to 17%. The control group voted at a rate of 54.5%, while the treatment group voted at a rate of 56.1%, so the estimated effect among the treated was 2.9% with a 2.8% standard error (see Table 1, column 1).

The results from the experiment may not appear impressive, but consider the estimate if the standard protocol had been used instead. Without the ability to roll unattempted subjects into the control group, the application rate falls from 55% to 19%. The low application rate inflates the standard error for the ultimate estimate from 2.8% to 6.3%—more than twice as large (see Table 1, column 2). Figure 1 displays the 95% confidence intervals for the rolling and standard protocols for the Boston experiment. Neither of the protocols generated biased estimates, but the rolling protocol offered statistical precision in the face of a 30% shortfall in volunteers.

[19]See Nickerson 2004a for a full description.

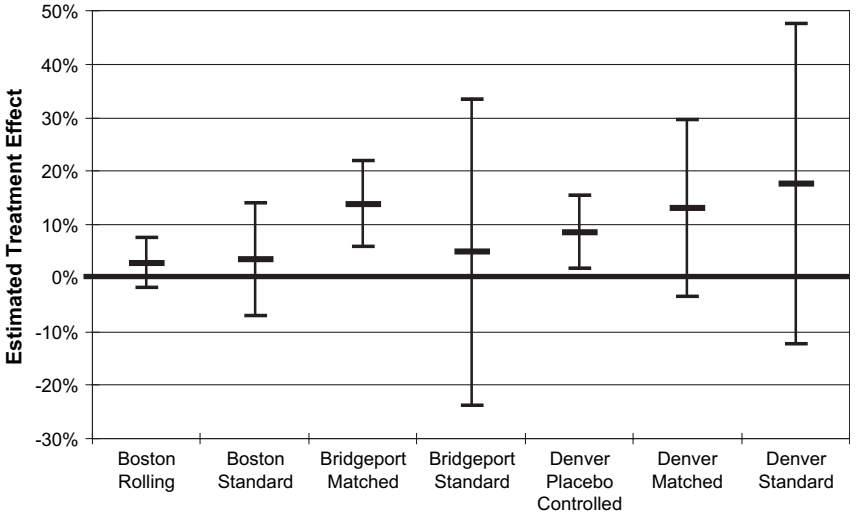

Fig. 1 Efficiency gained through the use of scalable protocols.

5 Matched Experimental Protocol[20]

It is not always possible to randomly order the application of treatment for subjects. The subjects may fall into discrete groups that cannot easily be integrated with each other. For instance, if a researcher wanted to measure the influence of participation in community building exercises on high school students' feelings of internal efficacy (with attending a school assembly serving as a control), it may not be possible to rank order the students because participation in the experiment depends upon a teacher agreeing to forgo class for the period. This hurdle does not necessitate the use of the standard experimental protocol. Instead of breaking the subjects into monolithic treatment and control groups, those groups can be subdivided by classroom. If a teacher decides to participate in the experiment, the students are divided randomly into treatment and control groups. Classrooms that were not permitted to be a part of the experiment are not included in the analysis and therefore do not adversely affect the application rate of the experiment.

More generally, matching treatment and control groups within the discrete units of analysis fixes the application rate and the treatment/control ratio, leaving on the size of the experiment vulnerable to problems in execution of the protocol. From the example above, if 100% of the teachers participate, the size of the experiment, N, remains the same. However, when only P classrooms are able to participate, then the experiment contains NP subjects. The variance for estimated treatment effect upon those treated using the matched protocol is

$$Var(\delta_{TOT})_M = \frac{\sigma^2}{A^2PNT(1-T)}. \tag{8}$$

[20]The matched protocol is often referred to as a "blocked" or "paired" design (for a more lengthy discussion see Wu and Hamada [2000, pp. 48–54]; Montgomery [2001, pp. 47–51, 126–140]; Riffenburgh [1998, pp. 18, 145, 284]; and Box et al. [1978, pp. 93–106]).

Once again as P decreases (i.e., problems increase or participation decreases) in Eq. (8) the variance of the estimate rises. Comparing Eq. (8) to Eqs. (6) and (7), the following proposition can be derived:

Proposition 4. *For* $0 < P < 1$*, the estimate of the treatment effect upon those treated using the matched protocol exhibits less variance (hence more statistical precision) than using the standard protocol but more variance than the rolling protocol, assuming* $\sigma^2_{matched} = \sigma^2_{standard}$*. (Proof in appendix)*

The intuition behind the proof is similar to the proof for Proposition 3. Whereas shortfalls in the standard protocol are squared, the matched protocol does not magnify problems. However, the rolling protocol does a slightly better job of hiding problems by shifting the burden to the treatment/control ratio.

A door-to-door voter mobilization experiment in Bridgeport, Connecticut, during 2001 presents an excellent illustration of the power of the matched protocol.[21] A nonprofit community group, ACORN, encouraged Bridgeport residents to vote during the month prior to a school board election. The voter rolls were obtained from the city clerk, and residents on each street were evenly divided into treatment and control groups.[22] Thus, each street served as the unit within which subjects were matched and randomly divided into the treatment conditions.

ACORN felt that it could successfully apply the treatment to most of the city during the weeks leading up to the election. These lofty ambitions were thwarted by two factors. First, the initial application rate was very low, so weekdays were spent rewalking the areas that were covered the prior weekend to raise the contact rate. The ultimate rate of application was a respectable 28%, but coverage of the city was significantly curtailed. Second, the mayoral race in a neighboring town was more competitive than the Bridgeport school board elections, so ACORN decided to shift resources away from Bridgeport to the mayoral race. As a result, only a small fraction of the anticipated labor was available for the Bridgeport experiment. The combination of these two factors meant that the experiment was one-tenth the size of the study initially planned and not as many streets were canvassed.

The resulting experiment using the matched protocol consisted of 1900 subjects living in neighborhoods canvassed by volunteers. Half of the subjects were assigned to the treatment group and contacted 28% of the time. The precision of the experiment was sufficient to detect the estimated 14% treatment effect with a standard error of 5% (see Table 1, column 3). In contrast, the standard protocol would have included 19,184 subjects, but a paltry 3% of the treatment group would have received the treatment. The estimated treatment effect, 5%, is dwarfed by a 17% standard error (see Table 1, column 4). The standard error for the standard protocol is 340% of the standard error of the matched protocol, and Fig. 1 displays the 95% confidence interval for each estimate. Even though both protocols yield unbiased results, the need for efficient design is obvious. Despite being only one-tenth the planned size, the Bridgeport experiment provided statistically significant results.

The flexibility of the matched protocol also deserves emphasis. The matching of the treatment and control groups can take place for large groups like schools or for pairs of

[21]See Gerber et al. (2003) for a full description.

[22]The rolling protocol is infeasible for door-to-door efforts because walking from house to house in a random order would be impractical.

individuals or organizations. The logic remains the same: if application of the treatment is attempted, then include both the treatment and matched control group into the analysis. It is difficult to conceive of a field experiment in which it is not possible to match subjects and divide these subgroups into treatment and control groups before conducting the experiment.[23]

The idea of matching subjects within social groups provides a good place to discuss violations of the stable unit value treatment assumption (SUVTA; see Rubin 1986). The experimental analysis discussed in this article assumes that each version of the treatment regime is represented (e.g., the drugs taken by the patients do not vary in effectiveness) and there is no cross-contamination or interference between subjects (e.g., treatment and control patients do not share drugs in any systematic manner). Registered voters residing on the same street are unlikely to interact sufficiently to bias inferences. In contrast, students within classrooms, as described in the beginning of the section, are extremely likely to interact with one another, so cross-contamination of treatment is a serious concern. The validity of SUVTA will depend upon the specific context of the experiment. In situations in which violations of SUVTA are likely, the researcher should not randomize individual subjects within the group but should instead consider using the groups themselves as the unit of randomization. Violations of SUVTA are not unique to the matching protocol, but blocking within social groups highlights the necessity for a researcher to pay careful attention to the proper unit of randomization and analysis to avoid biased inferences.

6 Placebo-Controlled Protocol

Both the rolling and matched protocols will protect the application rate from problems with the execution of the experiment, but offer little assistance in the face of extremely low baseline rates of application. In such instances, an ideal experimental protocol would not be dependent upon the application rate at all. Such experiments are common in the medical sciences where placebos, or dummy treatments, are used (see Riffenburgh 1998). Researchers randomly determine whether subjects receive an experimental drug or a sugar pill and then compare the outcomes of subjects who complete the prescribed regime. Often such medical experiments are "double-blind" meaning neither the patient nor the doctor prescribing the treatment knows whether the experimental drug or the placebo has been administered.

The same principle can be applied to social science research by conducting parallel treatment regimes in which one treatment is of substantive interest and the second is a placebo. Rather than rely upon a control group that receives no attempted treatment, the group receiving the placebo can serve as the baseline for comparison for the treatment group. Examining the two sets of subjects who complete the treatment and placebo regimes provides an unbiased estimate of the treatment effect, assuming that (1) the two treatments have identical compliance profiles; (2) the placebo does not affect the dependent variable; and (3) the same type of person drops out of the experiment for the two groups. These three assumptions are more onerous than either the matching or rolling protocols and will be discussed at the end of the section.

[23]For practical reasons of implementation, the randomization into treatment and control groups can occur prior to the subject's decision to participate in the experiment. However, it is important that the decision to participate has nothing to do with which treatment condition the subject was assigned to. Thus, neither the potential subject nor the researcher should be aware of the potential subject's assignment during the recruitment process.

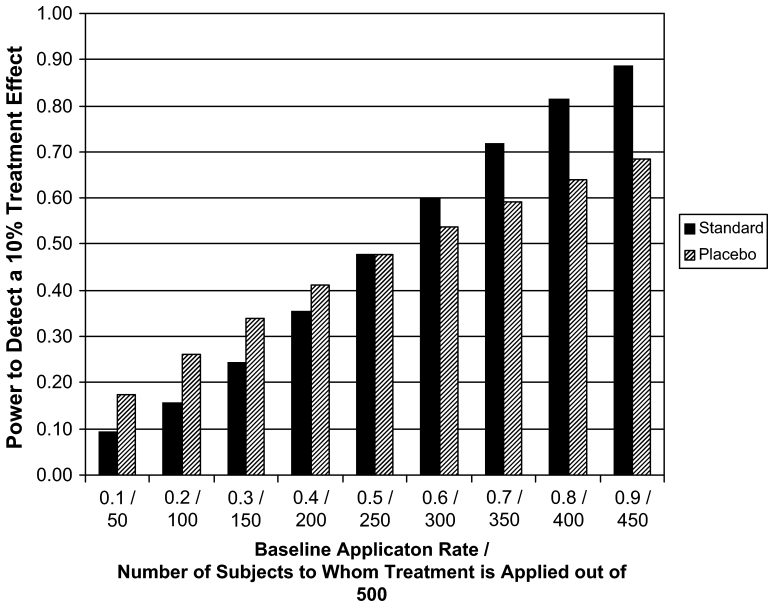

Fig. 2 Power comparison of placebo-controlled and standard experimental protocols for varying application rates and perfectly executed protocols.

The benefit of the placebo design is that only those subjects to whom an entire regime has been applied are considered.[24] Thus, the application rate is 100%. However, the number of subjects analyzed in the experiment declines considerably.[25] The trade-off between the number of subjects and the application rate leads to the following proposition:

Proposition 5. *Given an equal number of treated subjects*[26] *with identical variance in the dependent variable, when subjects are difficult to treat (i.e., application rates are low), the placebo-controlled protocol offers more statistical power than the standard, matched, or rolling protocols. (Proof in appendix)*

The intuition behind the proof is the same as Proposition 4. While the application rate is squared for protocols in which the control group remains untouched by the researcher, the application rate is irrelevant for the placebo-controlled protocol. However, conducting two efforts to treat subjects decreases the number of subjects in the experiment, so the placebo does not provide efficiency gains in all instances. Figure 2 compares the placebo-controlled protocol with the standard protocol's ability to detect a 10% treatment effect for a variety of application rates given 1000 subjects, a 50–50 split between treatment and control, an equal number of applications of the treatment, and a fully implemented protocol ($P = 1$). Notice that the two protocols have the same power when treatment/control ratio equals the application rate. This is true for all values of T and A, provided that each type of experiment contains an equal number of treated subjects.

[24]Designs in which subjects complete only a portion of the treatment and placebo regimes can also be considered.
[25]It should also be noted that subjects who did not complete the treatment or placebo regime can be discarded from the analysis only when all three of the placebo assumptions are true.
[26]This assumption is made as a proxy for the cost of conducting the experiment. In most field experiments the major expense lies in applying the treatment to subjects. In cases in which this is not the case, the relevant comparison between placebo and controlled protocols may not be with an equal number of treated subjects.

Comparing the placebo-controlled protocol's robustness to problems encountered in the field with the other protocols (standard, matched, and rolling) is difficult because the statistical power of the protocols depends upon the baseline application rate and the treatment/control ratio. However, the efficiency of the placebo-controlled protocol can be established within boundaries. Placebo-controlled protocols are efficient for much the same reason as the matched protocol. The treatment/control ratio is predetermined and the subjects included in the analysis have all received a treatment, so problems in the field are most likely to affect the number of subjects. The variance of the estimated treatment effect upon those treated for the placebo-controlled protocol can be expressed as

$$Var(\delta_{TOT})_C = \frac{\sigma^2}{PN_cT(1-T)}. \tag{9}$$

The notation N_c is used to signify that the population of subjects in the placebo-controlled protocol is not identical to the other three protocols discussed. From Eq. (9) it is possible to derive the following three conclusions:

Proposition 6. *Given an equal number of subjects receiving a treatment with equal variance in the dependent variable:*

(1) The placebo-controlled protocol generates a more efficient estimate of the treatment effect than the standard protocol when $T > PA$;

(2) The placebo-controlled protocol generates a more efficient estimate of the treatment effect than the matched protocol when $T > A$;

(3) The placebo-controlled protocol generates a more efficient estimate of the treatment effect than the rolling protocol when $A(1 - PT) < T(1 - T)$. (Proof in appendix)

The power of the placebo-controlled protocol is illustrated by a voter mobilization experiment in Denver conducted during the 2002 primaries. A total of 8311 households in the city were mapped into geographically clustered turfs of roughly 60 houses. Households in each one of these turfs were then randomly divided into one of three groups: control, encouragement to vote, and encouragement to recycle.[27] A group of environmental activists had agreed to canvass for the week leading up to the primaries. Unfortunately rain canceled two days of work, days with good weather had low participation rates, and volunteers who did arrive worked extremely inefficiently and for very few hours. Ultimately the campaign contacted 283 subjects in the voting group and 273 subjects in the recycling group for an overall application rate of 17%. The group contacted about voting did so at a rate of 48%, while the group contacted about recycling voted at a rate of only 39%, for an estimated treatment effect of 9% with a standard error of 4% (see Table 1, column 5).

The mapping of households into turfs also allowed for the matched protocol to be used comparing the control group to those encouraged to vote. While 5540 households were in the eligible sample,[28] less than a third of these neighborhoods were attempted, so the N for the matched protocol is 3398. The group that was to receive no contact from the experiment voted at a rate of 38%, while the group encouraged to vote did so 40% of the time. Factoring in the 17% application rate, the matched design estimated an effect among

[27]For more information on the experiment see Nickerson (2004b).

[28]One-third of the sample was assigned to receive the recycling message and would not be part of the matched experimental design.

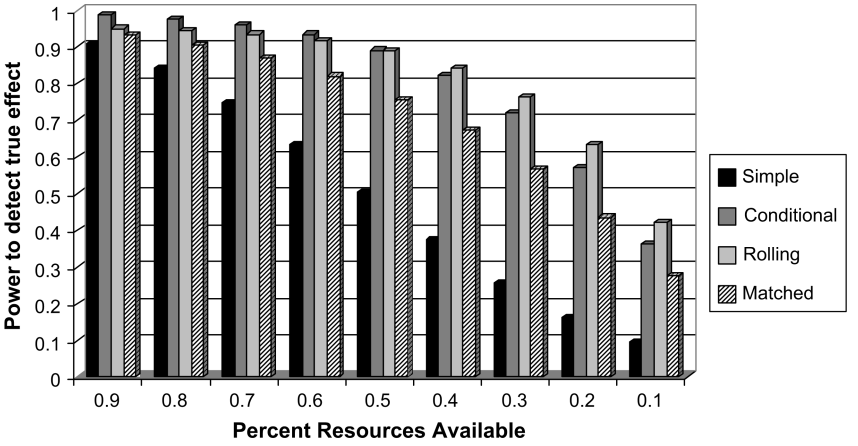

Fig. 3 Power comparison of simple, matched, rolling, and placebo-controlled protocols. The statistical power was calculated using $A = 0.33$, $N = 10000$, $T = 0.5$, and a one-tailed test with $\alpha = 0.05$.

the treated of 13.2% with a standard error of 10% (see Table 1, column 6). While the matched protocol does not have the precision of the placebo-controlled protocol, it is far more useful than the standard protocol, which has an application rate of 5% and a standard error of 18% (see Table 1, column 7). Figure 1 provides a graphical presentation of the confidence intervals for the three protocols. The extremely low rate of application made the placebo-controlled protocol much more powerful and robust to field problems than either the standard or matched protocols in this instance.

The assumptions behind the placebo-controlled protocol deserve special mention. Finding an appropriate placebo may not be easy for many topics in political behavior, since the two major requirements can be in tension. A placebo that has no independent effect upon outcome measures is unlikely to have a similar rate of application profile as the treatment. For instance, suppose a researcher seeks to determine how political documentaries change pre-existing beliefs. In addition to measuring attitudes before and after viewing the documentary, a randomly selected placebo group is set up to watch the romantic comedy *Shakespeare in Love*. While *Shakespeare in Love* is unlikely to change a subject's political beliefs, the set of subjects who will sit through the romantic comedy and the political documentary are likely to be very different. Thus, it is possible that differential rates of treatment application and attrition will result from the study. Unfortunately, showing another type of documentary (i.e., nature) may not solve the differential completion and attrition problem. Screening a different political documentary would be a poor placebo because attitudes might change as a result of the viewing. In short, conflicts can arise from the demand that placebos have no causal connection to outcome variables and have identical rates of completion as the treatments of interest. Placebos used in the social sciences must be chosen with great care, otherwise the results will be biased.

7 Discussion

Incorporating one of the efficient, scalable protocols increases statistical precision with little additional work on the part of the researcher. To illustrate the efficiency gains, suppose an experiment possesses 10,000 subjects divided equally between treatment and control groups with a baseline application rate of 33%. Figure 3 illustrates the statistical

power of each of the four protocols to detect a 10% treatment effect at various levels of funding (i.e., percent of protocol completed).[29] When an experiment receives 90% of the necessary funding, the differences in statistical power are small. However, as the funding level decreases the superiority of the scalable experimental protocols over the standard protocols becomes apparent. With only 30% of the necessary resources, the standard protocol can accurately reject the null hypothesis, $H_0 = 0$, only 26% of the time. In contrast, the matched protocol can do so 57% of the time, the rolling 75%, and the placebo 72%. In other words, when 70% of the anticipated resources are lacking, the power of the matched design is twice that of the standard protocol, and the rolling and placebo-controlled protocols are almost three times as powerful as the standard protocol. Implementing the rolling or matched protocols requires no additional work on the part of a researcher; with just a little advanced planning, any experiment conducted in the field should incorporate some type of scalable experimental protocol into the research design.

As noted earlier, successfully applying the treatment regime in voter mobilization experiments consists of successfully finding the person at the door. Thus, delivery, receipt, and adherence are collapsed into one parameter. For most topics in political behavior, the treatment regime will not be so simple, but the same principles should apply. Whether a researcher faces trouble with delivery, receipt, or adherence, the problems are best shifted to either the treatment-control ratio or the number of subjects. The definition of success for each facet of the treatment regime will necessarily differ across topics, but the logic and efficiency of the scalable designs still hold.[30]

This article focuses upon the rate of successful application of the treatment regime for two reasons. First, low application rates decrease statistical efficiency, which is important in its own right, but also allows the researcher to conserve resources to devote to other problems. Second, the application of the treatment is where many field experiments run aground. In their natural habitat, subjects are less prone to complying with treatment regimes than in a laboratory setting. Careful attention to protocol design can shift failure to treat issues away from the rate of application to less sensitive parameters such as treatment-control ratio and the number of subjects in the experiment. More complicated analytic techniques are not required to analyze the scalable results, but advanced planning is essential.

Both the placebo and matched protocols restrict the sample of subjects to be analyzed. Typically, restricting a sample raises concerns about external validity. However, the procedures described in this paper do nothing to diminish the external validity of the experimental findings, because the only subjects dropped from the analysis are those without whom the researcher could not even attempt the execution of the experiment. It is logically impossible to derive information from a sample of individuals upon which no experiment was attempted. The restricted samples in the matched and placebo protocols only make explicit the epistemological problem that previously existed. The validity of findings outside of the experimental population is always an open question; the problem is no worse under the scalable protocols than under the standard protocol.

Readers should be careful to note that problems applying the treatment regime, which do not necessarily introduce bias, may be indicative of larger problems with the execution of the experiment that can bias results. Failure to treat often results in, or is caused by, subject attrition, which may introduce bias to an experiment. The whole point of

[29]The apparent superiority of the placebo-controlled protocol over the rolling and matched protocols is due to the relatively low application rate of 33%.

[30]However, there may be instances in which the definition of successful compliance may be murky and subjective. See Heitjan (1999) for a medical example and discussion.

randomized experimentation is to avoid bias, so such problems should be the primary concern of researchers. Scalable protocols cannot solve bias introduced by subject attrition,[31] errors in randomization (i.e., the assignment of treatment is correlated with an outside factor, which may or may not be causal of the outcome variable), or measurement error. Preventing or correcting the bias may consume sufficient time and resources that a researcher may be forced to scale back the size of the experiment. In such instances, the scalable protocols are indirectly useful in addressing the concern of bias.

Scalable protocols have other advantages beyond improving efficiency. For instance, the rolling protocol can also avoid thorny ethical issues stemming from depriving subjects of the experimental treatment. Imagine that a school system is attempting to implement a curriculum based on new computers to be placed in the classroom. A randomized study of the effectiveness of the new curriculum and computers would be interesting, but placing students in a control group where they do not receive the new computers denies the children an educational tool. If the school system cannot afford to provide every classroom the necessary equipment immediately, the order in which classrooms receive the materials could be randomized. Classrooms that do not receive the materials by the end of the semester or year constitute a control group, but no students were deprived of the tools to learn because of the experiment. Without the experiment, the lack of educational equipment is simply a misfortune, but the rolling protocol can turn the pre-existing shortfall into a valuable source of information that can guide future administrative decisions (see Boruch et al. 2000, pp. 170–173, for brief and useful thoughts on the ethics of randomized experiments).

An alternative means of making experiments more precise is the use of control variables. Successful control variables are predictive of the outcome variable but are not affected by the treatment.[32] Including such variables in the analysis will not reduce the overall variance of the estimand, but the unexplained variance will be lessened. Collecting a handful of good control variables is often a cost-effective means of increasing the precision of an experiment (especially when compared with increasing sample size). There is no reason why the use of control variables cannot be combined with the scalable designs described in this paper. Indeed, the collection of control variables is one more facet of an experiment that could meet with problems and require additional attention, thereby necessitating the efficiency of scalable protocols.

Controlled experiments offer unbiased estimates of treatment effects and field experiments mitigate concerns about external validity. As with any study conducted in the field, problems will arise that compromise the statistical power of the experiment. Careful planning and the use of scalable experimental designs can mitigate the impact of problems encountered in the field. The execution of efficient experimental protocols may be slightly more difficult than the standard experimental protocol, but the increased precision of the estimates derived more than compensates.

Appendix

Proof of Proposition 1: Compare the experiment described by Eq. (1) to a second experiment that shares the population variance, σ, and a treatment/control ratio, T, but

[31] The placebo-controlled design could conceivably address the attrition problem. The assumption that is required is that identical types of subjects drop out of the experiment under both the treatment and placebo regimes. The assumption may be reasonable in many instances, but will prove unverifiable in most instances. Such analysis should be undertaken with great caution and transparency to let the reader judge the validity of the assumption.

[32] Background or demographic information is most often used.

differs with an application rate of AX, where $0 \leq X \leq 1$, and contains $N + n$ subjects. The variance of this second experiment can be expressed as

$$Var_{E2} = \frac{\sigma^2}{(AX)^2(N+n)T(1-T)}. \tag{A1}$$

Suppose $X < \sqrt{N/(N+n)}$. Then, $Var_{E2} = \sigma^2/(AX)^2(N+n)T(1-T) > \sigma^2/A^2NT(1-T) = Var_E$. That is, when the application rate declines by $\sqrt{N/(N+n)}$ the increase in variance offsets the decrease in variance from the additional n subjects. The inverse relationship holds for $X > \sqrt{N/(N+n)}$.

Proof of Proposition 2: Compare the experiment described by Eq. (1) to another experiment that shares the population variance, σ, and the number of subject, N, but differs with an application rate of AX, where $0 \leq X \leq 1$, and where $n \in \mathbb{Z}$ subjects are moved from the control to the treatment group. The variance for the first experiment can be expressed as in Eq. (1) and the variance of the second experiment can be expressed as

$$Var_{E3} = \frac{\sigma^2}{(AX)^2(t+n)\left(1-\frac{t+n}{N}\right)}. \tag{A2}$$

Suppose $X < \sqrt{t(N-t)/(t+n)(N-t-n)}$; then $Var_{E3} = \sigma^2/(AX)^2(t+n)(1-t+n/N) > \sigma^2/A^2t(1-t/N) = \sigma^2/A^2NT(1-T) = Var_E$. That is, when the application rate declines by more than $\sqrt{t(N-t)/(t+n)(N-t-n)}$ the increase in variance offsets any advantage generated by shifting n subjects from the control to the treatment group. The inverse relationship holds for $X > \sqrt{t(N-t)/(t+n)(N-t-n)}$.

Proof of Proposition 3: Rewrite Eq. (7) as

$$Var_R = \left(\frac{1-T}{P(1-PT)}\right)\frac{\sigma^2}{A^2NT(1-T)}. \tag{A3}$$

Note that $\lim_{T\to 0} 1 - T/P(1-PT) = 1/P$ and $\lim_{T\to 1} 1 - T/P(1-PT) = 0$. Thus, for $0 < P < 1$, $Var_R = (1 - T/P(1-PT))\sigma^2/A^2NT(1-T) < \sigma^2/P^2A^2NT(1-T) = Var_S$.

Therefore, for $0 < P < 1$, the rolling protocol estimator exhibits less variance (hence more statistical power) than the estimator for the standard protocol.

Proof of Proposition 4: For all $0 < P < 1$, $1/P < 1/P^2$. This implies that the following two conditions hold:

$$1) \quad Var_M = \frac{\sigma^2}{PA^2NT(1-T)} < \frac{\sigma^2}{P^2A^2NT(1-T)} = Var_S$$

$$2) \quad Var_R = \left(\frac{1-T}{P(1-PT)}\right)\frac{\sigma^2}{A^2NT(1-T)} < \frac{\sigma^2}{PA^2NT(1-T)} = Var_M,$$

assuming that σ^2 is the same for the smaller matched experiment as the population as a whole. Therefore, for $0 < P < 1$, the matched protocol exhibits less variance than the standard protocol but more variance than the rolling protocol.

Proof of Proposition 5: The variance for the placebo-controlled protocol may be expressed as

$$Var_C = \frac{\sigma^2}{aT(1-T)}, \tag{A4}$$

where a equals the number of subjects to whom a treatment could be applied. In the standard, rolling, and matched protocols, $a = ANT$. Substituting, we rewrite Eq. (A4) as

$$Var_C = \frac{\sigma^2}{ANT^2(1-T)}. \tag{A5}$$

When $T > A$, $\sigma^2/ANT^2(1-T) < \sigma^2/A^2NT(1-T) = Var_E$. Likewise, when $T < A$, $\sigma^2/ANT^2(1-T) > \sigma^2/A^2NT(1-T) = Var_E$. Thus, given an equal number of subjects to whom treatment is applied and $T > A$, the placebo-controlled protocol exhibits less variance than the standard, rolling, and matched protocols.

Proof of Proposition 6: Rewrite Eq. (9) as

$$Var_C = \frac{\sigma^2}{PaT(1-T)}, \tag{A6}$$

where a equals the numbers of subjects to whom a treatment is applied. In the standard, rolling, and matched protocols, $a = ANT$. Replacing the variance for the placebo-controlled estimator is:

$$Var_C = \frac{\sigma^2}{PANT^2(1-T)}. \tag{A7}$$

Suppose $T > AP$; then

$$Var_C = \frac{\sigma^2}{PANT^2(1-T)} < \frac{\sigma^2}{P^2A^2NT(1-T)} = Var_S.$$

Suppose $T > A$; then

$$Var_C = \frac{\sigma^2}{PANT^2(1-T)} < \frac{\sigma^2}{PA^2NT(1-T)} = Var_M.$$

Suppose $PT > P(1-PT)A/(1-T)$; then

$$Var_C = \frac{\sigma^2}{PANT^2(1-T)} < \left(\frac{1-T}{P(1-PT)}\right)\frac{\sigma^2}{A^2NT(1-T)} = Var_R.$$

References

Adams, William C., and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44:389–395.

Allison, Paul D. 2002. *Missing Data.* Thousand Oaks, CA: Sage.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996a. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996b. "Identification of Causal Effects Using Instrumental Variables: Rejoinder." *Journal of the American Statistical Association* 91:468–472.

Boruch, Robert, Brook Snyder, and Dorothy DeMoya. 2000. "The Importance of Randomized Field Trials." *Crime and Delinquency* 46:156–180.

Box, George E. P., William G. Hunter, and J. Stuart Hunter. 1978. *Statistics for Experiments.* New York: Wiley-Interscience.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand-McNally.

Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50:154–165.

Fisher, Ronald A. 1935. *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94:653–663.

Gerber, Alan S., Donald P. Green, and David W. Nickerson. 2003. "Getting Out the Vote in a Local Election: Results from Six Door-to-Door Canvassing Experiments." *Journal of Politics* 65:1083–1096.

Goodman, Jodi S., and Terry C. Blum. 1996. "Assessing the Non-random Sampling Effects of Subject Attrition in Longitudinal Research." *Journal of Management* 22:627–652.

Groves, Robert M. 1989. *Survey Errors and Survey Costs.* New York: Wiley.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Randomized Social Experiments." *Journal of Economic Perspectives* 9:85–110.

Heitjan, Daniel F. 1999. "Causal Inference in a Clinical Trial: A Comparative Example." *Controlled Clinical Trials* 20:309–318.

Kish, Leslie. 1965. *Survey Sampling.* New York: Wiley.

Lichstein, Kenneth L, Brant W. Riedel, and R. Grieve. 1994. "Fair Tests of Clinical Trials: A Treatment Implementation Model." *Advances in Behavior Research and Therapy* 16:1–29.

Little, Roderick J. A., and Donald B. Rubin. 1987. *Data Analysis with Missing Data.* New York: Wiley.

Miller, Roy E., David A. Bositis, and Denise L. Baer. 1981. "Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman." *International Political Science Review* 2:445–460.

Montgomery, Douglas C. 2001. *Design and Analysis of Experiments,* 5th ed. New York: Wiley.

Nickerson, David W. 2004a. "Volunteer Phone Calls Can Increase Turnout." Unpublished manuscript.

Nickerson, David W. 2004b. "Is Voting Contagious?" Unpublished manuscript.

Riffenburgh, R. H. 1998. *Statistics in Medicine.* San Diego: Academic.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

Rubin, Donald B. 1986. "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers?" *Journal of the American Statistical Association* 81:961–962.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* New York: Houghton Mifflin.

Wu, Chien-Fu, and Michael Hamada. 2000. *Experiments: Planning, Analyzing, and Parameter Design Optimization.* New York: Wiley-Interscience.