Article

Evaluation Review
2018, Vol. 42(4) 391-422

DOI: 10.1177/0193841X18799128
journals.sagepub.com/home/erx
SAGE

Optimal Allocation of Interviews to Baseline and Endline Surveys in Place-Based Randomized Trials and Quasi-Experiments

Donald P. Green[1], Winston Lin[2], and Claudia Gerber[3]

Abstract

Background: Many place-based randomized trials and quasi-experiments use a pair of cross-section surveys, rather than panel surveys, to estimate the average treatment effect of an intervention. In these studies, a random sample of individuals in each geographic cluster is selected for a baseline (preintervention) survey, and an independent random sample is selected for an endline (postintervention) survey. **Objective:** This design raises the question, given a fixed budget, how should a researcher allocate resources between the baseline and endline surveys to maximize the precision of the estimated average treatment effect? **Results:** We formalize this allocation problem and show that although the optimal share of interviews allocated

[1] Department of Political Science, Columbia University, New York, NY, USA
[2] Department of Statistics and Data Science, Yale University, New Haven, CT, USA
[3] International Monetary Fund, Washington, DC, USA

Corresponding Author:
Winston Lin, Department of Statistics and Data Science, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA.
Email: winston.lin@yale.edu

to the baseline survey is always less than one-half, it is an increasing function of the total number of interviews per cluster, the cluster-level correlation between the baseline measure and the endline outcome, and the intracluster correlation coefficient. An example using multicountry survey data from Africa illustrates how the optimal allocation formulas can be combined with data to inform decisions at the planning stage. Another example uses data from a digital political advertising experiment in Texas to explore how precision would have varied with alternative allocations.

Keywords

place-randomized trial, cluster-randomized experiment, quasi-experiment, repeated cross-section surveys, sample allocation

Surveys are widely used to measure outcomes in randomized control trials (RCTs) and quasi-experiments. Although only endline (posttreatment) outcome data are required for the estimation of treatment effects in RCTs, baseline (pretreatment) survey data may be helpful for improving statistical precision and power. In panel surveys, a common set of respondents is tracked over time from baseline to endline, allowing researchers to assess how the trajectories of individual subjects' outcomes in the treatment group compare with those of the control group. Optimizing the design of panel surveys for efficient estimation of average treatment effects (ATEs) has attracted increasing scholarly attention (McKenzie, 2012).

As Gail, Mark, Carroll, Green, and Pee (1996) discuss, panel surveys have important strengths and are often desirable for statistical precision, but they can also have important drawbacks in some contexts. Maintaining contact with baseline respondents may be costly or difficult, especially when tracking subjects who frequently change address or phone number (Parker & Teruel, 2005). A further concern is that the baseline interview may prime subjects in ways that alter their reaction to the treatment, distort their posttreatment survey responses, or cause nonresponse rates in the endline survey to differ between treatment and control groups (Flay & Collins, 2005; Solomon, 1949).

When treatments are administered to a set of geographic clusters (Boruch, 2005; Gail, Mark, Carroll, Green, & Pee, 1996), an alternative measurement design is to interview a random sample of individuals within each cluster at baseline and another random sample at endline. When researchers gather survey data using this repeated cross-section design with clusters of equal size, the ATE of the intervention may be estimated by comparing the average outcomes of treatment and control group clusters in the endline survey, adjusting for preexisting differences in the baseline survey.

A wide array of applications have used this design. Table 1 presents illustrative examples of repeated cross-section designs from a variety of substantive domains. For example, Ter Kuile et al. (2003) assessed the effects of bed nets on malaria among young children by randomly assigning 60 Kenyan villages to treatment and control. Random samples of children in each village were given medical exams at baseline, and new random samples were examined at endline. Another example is Gerber, Gimpel, Green, and Shaw (2011), which assessed the persuasive effects of political advertisements across 18 television markets by conducting a baseline survey within each market before the advertising campaign and drawing new samples within each market for the endline surveys. Indeed, the use of this design is common among experiments that assess the persuasive effects of political advertising, where automated phone surveys are conducted with distinct random samples of registered voters during baseline and endline periods. These automated surveys are directed at landline phone numbers associated with a particular address rather than a specific person, which makes it impractical to conduct panel surveys that track the same respondents over time. One of the empirical applications described below (Turitto, Green, Stobie, & Tranter, 2014) uses this design to assess the effects of digital advertising on behalf of a candidate for lieutenant governor of Texas. Although such studies are common, political campaigns rarely make the results public.

When using the repeated cross-section design to estimate the ATE of an intervention, a resource allocation question arises: In order to maximize the precision of the estimated ATE, how much of the survey budget should be allocated to the baseline survey as opposed to the endline survey? To our knowledge, none of the studies listed in Table 1 discuss this allocation problem.

This article begins by formalizing the allocation problem in a balanced experimental design (where equal numbers of clusters are assigned to treatment and control) and deriving a result that expresses the optimal allocation as a function of the budgeted number of survey interviews per cluster, the cluster-level correlation between the baseline measure and the endline outcome, and the intracluster correlation coefficient (ICC). We then show how insights from the formal analysis can be applied in practice, using data from the Afrobarometer surveys (Afrobarometer, 2009, 2015) for an illustrative example. Next, we discuss survey allocation in an imbalanced design,

Table 1. Examples of Place-Based Evaluations Using Repeated Cross-Section Surveys.

Study	Field/Topic	Summary and Main Findings	RCT or Quasi-Experiment	Baseline Survey	Endline Survey(s)
Bloom and Riccio (2005)	Jobs and public housing	Evaluates an employment initiative within public housing developments, implemented in six U.S. cities. Finds positive effects on earnings and employment for most housing projects that implemented the program correctly. The effects did not spark changes in overall social conditions or quality of life	RCT • Random assignment of 16 housing developments (6 treated, 10 control)	• *N* = 2,123 for treatment group • *N* = 2,651 for control group	• Follow-up 5 years later in 2003 • *N* = 2,700–4,500 (300–500 per housing project)
Gerber et al. (2011)	Politics	Explores the impact of political radio and television advertising on public opinion among registered voters in Texas. Finds ephemeral effects on voting preferences	RCT • Random assignment of 18 designated media markets to varying quantities of TV and radio ads	• Conducted a few days before the launch of the media campaign • *N* = 150 per media market (*N* total = 2,998)	• Two follow-ups: a week after intervention and second round 5 weeks later • *N* = 350 per week per media market (*N* = 7,022 for week 1) • Number of surveys and survey responses vary by week
Ter Kuile et al. (2003)	Public Health: Malaria	Studies the impact of insecticide-treated bed nets on malaria-associated morbidity in children under age 3 in Kenya. The nets reduced morbidity and improved weight gain in the treatment group	RCT • Randomly allocated 27 of 60 villages to treatment	• *N* = 889 (across 27 randomly selected villages out of 60)	• Two rounds (14 and 22 months after intervention) • *N* = 980 in survey (1) • *N* = 910 in survey (2) • Breakdown of treatment versus control interviews is unclear

(*continued*)

Table 1. (continued)

Study	Field/Topic	Summary and Main Findings	RCT or Quasi-Experiment	Baseline Survey	Endline Survey(s)
Smith, Ping, Merli, and Hereward (1997)	Public health: Contraception	Studies the impact of a revised and holistic contraception program in China on a range of outcome indicators, such as data quality on births and infant mortality. The results are mixed	RCT (overlapping surveys) ● Random assignment of 24 townships	● Control: $N =$ 8,603	● Five randomly selected townships (11,759 interviews; 2,676 of these respondents were also interviewed at baseline) ● Four townships assigned to treatment condition
Cheadle et al. (1995)	Public Health: Nutrition	Examines differences between evaluation tools when studying community-based nutrition programs. Identifies the "environmental indicator" as a good and low-cost evaluation tool compared with individual-level telephone and grocery surveys	RCT for two communities; quasi-experiment for 1 community ● Three intervention communities and seven control communities	● Random sample of stores from community clusters (15 stores per community) ● Phone survey of individuals, $N =$ 500 per community	● Two follow-ups after 2 years (1990: 21 stores per community; 1992: 26 stores per community) ● Phone survey of individuals, $N = 500$ per community
Murray et al. (1994)	Public Health: Cardiovascular disease	Investigates the impact of a 5–6 year heart health program in Minnesota on heart disease incidence, morbidity, and mortality. Mixed results	Quasi-experiment ● Nonrandom assignment of three communities to treatment and three to control	● $N = 300–500$ per community	● After 2 years (half of cohort) ● 4 years (other half) and both after 7 years ● $N = 300–500$ per community

(continued)

Table 1. (continued)

Study	Field/Topic	Summary and Main Findings	RCT or Quasi-Experiment	Baseline Survey	Endline Survey(s)
Farquhar et al. (1985)	Public Health: Cardiovascular disease	Describes the research design for a long-term field study to assess the impact of community health education in California for the prevention of cardiovascular disease	Quasi-experiment Nonrandom allocation	- Five communities ($N = 625$ per community)	- Two rounds (after end of campaign, and 3 years later)
Green, Wilke, Cooper, and Baltes (2016)	Social attitudes	Investigates the effects of exposure to video vignettes dramatizing the issues of violence against women, teacher absenteeism, and abortion stigma	Randomly allocated 28 rural trading centers to different messages	- 1,107 surveys in 28 trading centers in Uganda	- Follow-up surveys 2 months after videos were screened

Note. RCT = randomized control trial.

where the expense associated with administering treatment leads researchers to assign more clusters to control than to treatment. In the concluding section, we summarize the main lessons and discuss possible extensions to address a wider range of design considerations.

Model and Notation

To keep the allocation problem tractable, we will make a number of simplifying assumptions. First, suppose that we are planning an experiment or quasi-experiment with J clusters and that we are willing to assume the clusters are randomly assigned to treatment or control—either because the study is in fact a cluster-randomized experiment or because we believe the treated and untreated clusters are similar enough that modeling treatment as cluster randomized is reasonable. (In many nonrandomized studies, this assumption is *not* reasonable, and our analysis would need to be extended to consider possible roles for baseline covariates in reducing bias.)

Assume that baseline and endline interviews are equally costly and that our survey budget allows a total of S interviews.[1] One option is to allocate the entire budget to the endline survey, since treatment effects can be estimated without baseline data. Can precision be improved by allocating some interviews to a baseline survey and using the baseline data for blocking or covariate adjustment? If so, how many baseline and endline interviews should be conducted?[2]

For now, we assume a balanced design in which $J/2$ clusters are assigned to treatment and $J/2$ to control; the main ideas carry over to the case of an imbalanced design, which we discuss later. We also assume that any attempt to use a baseline covariate to improve precision will be done via linear regression adjustment, not blocking.[3] However, we do not assume that the true relationship between the outcome and the covariate is linear.[4]

Our analysis assumes that the J clusters were randomly selected from a much larger superpopulation and that the goal is to estimate an ATE (defined below) in the superpopulation. In practice, many studies use clusters that are not randomly drawn from any superpopulation. Some researchers therefore prefer a "finite population" framework in which statistical inferences are limited to the actual clusters in the study. Others defend the superpopulation framework on the grounds that it is useful to make inferences about "a hypothetical infinite population, of which the actual data are regarded as constituting a random sample" (Fisher, 1922, p. 311). However, the two frameworks tend to yield similar or even identical results, and the superpopulation framework often makes the mathematics easier. For a helpful discussion, see Reichardt and Gollob (1999).

Suppose each cluster j has a population of N_j individuals. Let y_{ij} and x_{ij} denote the endline outcome and the baseline covariate, respectively, for individual i in cluster j. Using the potential outcomes framework (Holland, 1986; Neyman, 1923; Rubin, 1974), let $y_{ij}(1)$ and $y_{ij}(0)$ denote the values that y_{ij} would take if cluster j were assigned to treatment or control, respectively. Averaging at the cluster level, let $Y_j = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}$, $X_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij}$, $Y_j(1) = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}(1)$, and $Y_j(0) = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}(0)$. Assume that our goal is to estimate the ATE in the superpopulation of clusters, weighting each cluster equally: $\text{ATE} = E[Y_j(1) - Y_j(0)]$. (Since each cluster j is randomly drawn from the superpopulation, the expectation in this formula is just the average over all clusters in the superpopulation.)

In each cluster j, the endline survey collects outcome data from a random sample E_j of n_{post} individuals, and the baseline survey (if conducted) collects covariate data from an independent random sample B_j of n_{pre} individuals.[5] Thus, the cluster has sample mean outcome $\widehat{Y}_j = \frac{1}{n_{\text{post}}}\sum_{i \in E_j} y_{ij}$ and sample mean covariate value $\widehat{X}_j = \frac{1}{n_{\text{pre}}}\sum_{i \in B_j} x_{ij}$. For simplicity, we assume the sample sizes n_{post} and n_{pre} are constant across clusters and are small relative to each cluster's population size N_j.

If $n_{\text{pre}} = 0$ (i.e., no baseline interviews are conducted), we will estimate the ATE using the unadjusted treatment–control difference in mean outcomes, weighting each cluster equally. Letting T_j equal 1 if cluster j is assigned to treatment and 0 otherwise, this estimator is given by:

$$\widehat{\text{ATE}}_{\text{unadj}} = \frac{1}{J/2}\sum_{j:T_j=1} \widehat{Y}_j - \frac{1}{J/2}\sum_{j:T_j=0} \widehat{Y}_j. \tag{1}$$

If $n_{\text{pre}} > 0$, we will use the estimated coefficient on T_j in an ordinary least squares regression of $\widehat{Y}_j$ on T_j and $\widehat{X}_j$. Let $\widehat{\text{ATE}}_{\text{adj}}$ denote the regression-adjusted estimator.[6]

The allocation problem is to choose n_{pre} and n_{post} to minimize the variance of the estimated ATE, subject to the constraint that $n_{\text{pre}} + n_{\text{post}} = n$, where $n = S/J$ is the budgeted number of survey interviews per cluster. Equivalently, the problem is to choose the proportion of baseline interviews $\pi = n_{\text{pre}}/n$.

To simplify the derivations and formulas, we will analyze the variances of $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{unadj}}$ when the treatment effect is homogeneous: Assume there is a constant τ such that $y_{ij}(1) - y_{ij}(0) = \tau$ for all i and j. Relaxing this assumption would complicate the analysis but would not necessarily be useful for study design, since the more complex formulas would involve quantities that are difficult to guess at the planning stage (such as the effect of treatment on the correlation between the covariate and the outcome).

We now define several quantities that affect the variance of the estimated treatment effect. The *between-cluster variance* of the potential outcomes is the variance of $Y_j(0)$ (or, equivalently, the variance of $Y_j(1)$, since we are assuming a homogeneous treatment effect):

$$\sigma^2_{y(\text{between})} = \text{Var}(Y_j(0)) = E\left[(Y_j(0) - \bar{Y}(0))^2\right], \tag{2}$$

where $\bar{Y}(0) = E(Y_j(0))$ and the expectations in these formulas are again just averages over all clusters in the superpopulation. The *average within-cluster variance* is given by:

$$\sigma^2_{y(\text{within})} = E\left[\frac{1}{N_j}\sum_{i=1}^{N_j}(y_{ij}(0) - Y_j(0))^2\right]. \tag{3}$$

Define the covariate's between-cluster variance $\sigma^2_{x(\text{between})}$ and average within-cluster variance $\sigma^2_{x(\text{within})}$ analogously. Assume $\sigma^2_{y(\text{between})}$, $\sigma^2_{y(\text{within})}$, $\sigma^2_{x(\text{between})}$, and $\sigma^2_{x(\text{within})}$ are all nonzero (as would be expected in most applications).

The ICC of the potential outcomes is given by:

$$\text{ICC}_y = \frac{\sigma^2_{y(\text{between})}}{\sigma^2_{y(\text{between})} + \sigma^2_{y(\text{within})}}. \tag{4}$$

We define ICC_x analogously and assume that $\text{ICC}_x = \text{ICC}_y$ (which may be a reasonable approximation if the covariate is a baseline version of the outcome). In what follows, it will be convenient to work with the quantity

$$K = \frac{1}{\text{ICC}_y} - 1 = \frac{\sigma^2_{y(\text{within})}}{\sigma^2_{y(\text{between})}}. \tag{5}$$

The *between-cluster correlation* between the covariate and each potential outcome is the correlation between X_j and $Y_j(0)$ (or, equivalently, the correlation between X_j and $Y_j(1)$):

$$\rho = \frac{\text{Cov}(X_j, Y_j(0))}{\sigma_{x(\text{between})}\sigma_{y(\text{between})}} = \frac{E[(X_j - \bar{X})(Y_j(0) - \bar{Y}(0))]}{\sigma_{x(\text{between})}\sigma_{y(\text{between})}}. \tag{6}$$

Equivalently, ρ is the square root of the R^2 that would be obtained if we could run a regression of $Y_j(0)$ on X_j in the superpopulation of clusters.

Results for Balanced Designs

The Appendix shows that the variance of $\widehat{\text{ATE}}_{\text{unadj}}$ is approximately

$$\frac{4}{J}\sigma^2_{y(\text{between})}(1+\frac{K}{n_{\text{post}}}), \tag{7}$$

while, for large enough J, the variance of $\widehat{\text{ATE}}_{\text{adj}}$ is approximately

$$\frac{4}{J}\sigma^2_{y(\text{between})}(1+\frac{K}{n_{\text{post}}})\left[1-\rho^2(1+\frac{K}{n_{\text{pre}}})^{-1}(1+\frac{K}{n_{\text{post}}})^{-1}\right]. \tag{8}$$

The factor $\frac{4}{J}\sigma^2_{y(\text{between})}$ is what the variance of the treatment–control difference in mean outcomes (weighting each cluster equally) would be if we could observe each cluster's population mean outcome Y_j. The next factor, $(1+K/n_{\text{post}})$, inflates the variance because each cluster's sample mean outcome $\widehat{Y}_j$ is a noisy estimate of Y_j. Finally, linear regression adjustment improves asymptotic precision, multiplying the variance by a factor of approximately $[1-\rho^2(1+K/n_{\text{pre}})^{-1}(1+K/n_{\text{post}})^{-1}]$, in which the squared correlation ρ^2 between the population means X_j and $Y_j(0)$ is attenuated by $(1+K/n_{\text{pre}})^{-1}(1+K/n_{\text{post}})^{-1}$ because the sample means $\widehat{X}_j$ and $\widehat{Y}_j$ are noisy estimates of X_j and Y_j.

The approximation to the variance of $\widehat{\text{ATE}}_{\text{adj}}$ may be improved by multiplying formula (8) by the degrees-of-freedom correction factor $(J-3)/(J-4)$ (Cox & McCullagh, 1982, p. 547). This factor is close to 1 when J is large.

The optimal allocation of interviews between the baseline and endline surveys is derived in the Appendix. If $|\rho| \leq K/n$, allocating all interviews to the endline survey is optimal (unless baseline interviews are desired for reasons other than improving precision). On the other hand, if $|\rho| > K/n$, the proportion of baseline interviews π that minimizes the approximate variance of $\widehat{\text{ATE}}_{\text{adj}}$ is

$$\pi_{\text{opt}} = 1-\frac{1+K/n}{1+|\rho|} > 0, \tag{9}$$

and, for large enough J, allocating $\pi=\pi_{\text{opt}}$ and using $\widehat{\text{ATE}}_{\text{adj}}$ is more efficient than allocating all interviews to the endline survey and using $\widehat{\text{ATE}}_{\text{unadj}}$.

To interpret formula (9) and its requirement that $|\rho| > K/n$, note that:

- $\pi_{\text{opt}} < 0.5$ (since $K > 0$ and $|\rho| \leq 1$). Thus, the proportion of interviews allocated to the baseline survey should always be less than one-half.
- π_{opt} is an increasing function of $|\rho|$ and n and a decreasing function of $K = 1/\text{ICC}_y - 1$. Intuitively, the usefulness of collecting data on baseline covariates depends on both $|\rho|$ (the strength of the true correlation between population mean covariate values and population mean potential outcomes at the cluster level) and the signal-to-noise ratio in the sample means (which improves with larger values of n and the ICC).[7]
- Related to the previous point, the condition $|\rho| > K/n$ is needed in order for π_{opt} in formula (9) to be positive. Otherwise, the true covariate–outcome correlation ρ is not strong enough, relative to the measurement error in the sample means, to make it worthwhile to allocate any interviews to the baseline survey.
- For any given values of ρ and K, as n goes to infinity, π_{opt} approaches an upper limit of $|\rho|/(1 + |\rho|)$.

Example Using Afrobarometer Data

When deciding how to allocate a survey budget between baseline and endline interviews, one does not know the values of ρ and the ICC, but it may be possible to form educated guesses using external data. This example illustrates the types of calculations involved, using data from the Afrobarometer, an ongoing series of cross-section public opinion surveys on democracy, governance, economic conditions, and related issues in African countries (Afrobarometer, 2009, 2015).

The first round (wave) of Afrobarometer surveys was conducted in 12 countries from 1999 to 2001. More recent rounds have included over 35 countries, with representative samples of 1,200 or 2,400 noninstitutionalized adult citizens in each country. The Afrobarometer surveys are useful for illustrative purposes given the large number of randomized trials conducted in Africa that use surveys to measure outcomes, the large number of respondents in each country at each point in time, and the wide array of outcomes measured (which allows us to consider outcomes with different ICCs and different values of ρ).

In order to simulate the country-level assignment typical of many quasi-experiments that estimate the effects of national policies on outcomes (e.g., Dorn, Fischer, Kirchgässner, & Sousa-Poza, 2007; Welsch, 2007), we use data from the 20 countries that were included in both the fourth (March

2008 to June 2009) and the fifth (October 2011 to September 2013) rounds of Afrobarometer surveys.[8] We focus on two outcome variables:

- Economic optimism: "Looking ahead, do you expect the following to be better or worse: Economic conditions in this country in 12 months' time?" (coded on a scale of 1 = "*much worse*" to 5 = "*much better*").[9]
- Inclination to protest: "Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: Attended a demonstration or protest march?" (coded on a scale of 0 = "*no, would never do this*" to 4 = "*yes, often*").[10]

Consider the problem of allocating a survey budget in a cluster-randomized experiment or quasi-experiment where the main outcomes of interest resemble the economic optimism and protest inclination variables. Suppose it has already been decided that the experiment will include 20 clusters, with 10 clusters assigned to treatment and 10 to control, and the budget allows a total of n interviews per cluster. For illustrative purposes, we will show calculations for both $n = 100$ and $n = 500$. (The larger sample size is similar to those in several of the evaluations listed in Table 1, such as Gerber et al. 2011.) Should a baseline survey be fielded, and if so, how should the interviews be allocated between the baseline and endline surveys?

To apply formula (9), we have the number of interviews per cluster n, but we need to estimate $K = 1/\mathrm{ICC}_y - 1$ and ρ for the main outcomes. If the interval between the proposed baseline and endline surveys is approximately the same as that between the fourth and fifth rounds of Afrobarometer surveys, we can use the Afrobarometer data to estimate both ICC_y and ρ.

Here, we use the analysis of variance (ANOVA) estimator of ICC (Donner, 1986, p. 68; Ridout, Demétrio, & Firth, 1999, p. 138). The estimated ICCs for economic optimism are 0.180 and 0.231 in the fourth and fifth rounds of the survey, while for inclination to protest, the corresponding estimates are 0.0425 and 0.0458. These translate into estimates for K of 4.56 or 3.33 (economic optimism) and 22.5 or 20.8 (inclination to protest).

The simplest way to estimate ρ is to just use the observed correlation between the fourth- and fifth-round country-level means of the relevant variable. These correlations are 0.578 for economic optimism and 0.681

for inclination to protest. However, ρ in formula (9) is the correlation between the cluster-level *population* means of the covariate and outcome in the absence of treatment, while the observed correlation ρ_{obs} between the *sample* means is expected to be somewhat attenuated (because the sample means are noisy estimates of the population means). A more refined estimate of ρ (derived in the Appendix) is

$$\widehat{\rho} = \rho_{\text{obs}}\sqrt{\left(1 + \frac{K_4}{n_4})(1 + \frac{K_5}{n_5}\right)}, \tag{10}$$

where $n_4 = 849.2$ and $n_5 = 1,100.0$ are the harmonic means of the country-level sample sizes in the fourth- and fifth-round surveys, and K_4 and K_5 are the estimates of K given above. Using this method, we obtain $\widehat{\rho} = 1.004 \times \rho_{\text{obs}} = 0.581$ for economic optimism and $\widehat{\rho} = 1.023 \times \rho_{\text{obs}} = 0.696$ for inclination to protest. The refinement does not matter much in this example, but it can matter when the sample sizes in the external data source are smaller or the ICCs are smaller. For example, without changing the ICCs, if we had $n_4 = n_5 = 100$, formula (10) would yield $\widehat{\rho} = 1.22 \times \rho_{\text{obs}}$ for inclination to protest.

The boundary condition $|\rho| > K/n$ is easily satisfied given our planned number of interviews per cluster ($n = 100$ or $n = 500$) and any of the above estimates of ρ and K, so we can use formula (9) to calculate π_{opt}, the optimal share of interviews to allocate to the baseline survey. The fourth- and fifth-round survey estimates of K are close enough that the choice between them hardly makes a difference; we use the fifth-round estimates in the remainder of this example. If $n = 500$, formula (9) yields $\pi_{\text{opt}} = 0.36$ for economic optimism and $\pi_{\text{opt}} = 0.39$ for inclination to protest, while if $n = 100$, the same formula yields $\pi_{\text{opt}} = 0.35$ for economic optimism and $\pi_{\text{opt}} = 0.29$ for inclination to protest.[11] As shown below, the precision of the estimated ATEs does not change dramatically as the proportion of baseline interviews π varies between 36% and 39% (for $n = 500$) or between 29% and 35% (for $n = 100$), so whether π is optimized for one outcome variable or the other (or a compromise between them) will not matter much in this example.

Figure 1 shows how the proportion of baseline interviews π affects the standard error (SE) of the estimated ATE on economic optimism:

- Near the top left corner, the two points marked with a filled triangle (for $n = 100$) and circle (for $n = 500$) show the SE of the unadjusted estimate $\widehat{\text{ATE}}_{\text{unadj}}$ when all interviews are allocated to the endline survey. To compute these SEs (0.275 and 0.272), we take the square

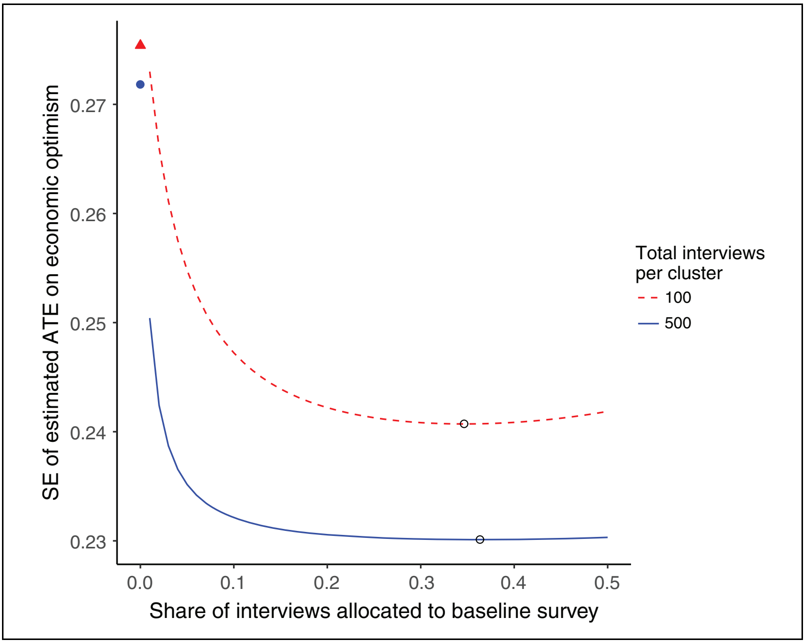

Figure 1. Survey allocation and precision when the outcome variable is economic optimism. Near the top left corner, the filled triangle (for $n = 100$ interviews per cluster) and circle (for $n = 500$) show the standard error (SE) of the unadjusted estimate of average treatment effect when all interviews are allocated to the endline survey. The curves plot the SE of the regression-adjusted estimate against the share of interviews allocated to the baseline survey. The open circle on each curve marks the optimal baseline share. See text for details.

root of formula (7) with $J = 20$, $K = 3.33$, and $\sigma^2_{y(\text{between})}$ set to 0.367, the unbiased estimate from the ANOVA (Donner, 1986, p. 68) applied to the fifth-round survey data.

- The two curves (dashed for $n = 100$ and solid for $n = 500$) show the approximate SE of the regression-adjusted estimate $\widehat{\text{ATE}}_{\text{adj}}$, calculated by multiplying the asymptotic variance from formula (8) by the degrees-of-freedom correction $(J - 3)/(J - 4)$ and then taking the square root, with $n_{\text{pre}} = \pi n$, $n_{\text{post}} = (1 - \pi)n$, $\rho = 0.581$, and the same values as above for J, K, and $\sigma^2_{y(\text{between})}$.
- The open circle on each curve marks the optimal allocation from formula (9). The optimal baseline shares are $\pi_{\text{opt}} = 0.346$ (for

> $n = 100$) and $\pi_{\text{opt}} = 0.363$ (for $n = 500$), achieving SEs of 0.241 and 0.230, respectively. However, both curves are relatively flat over a wide range of allocations: Virtually the same SEs could be achieved by allocating anywhere from 20% to 50% of the interviews to the baseline survey. Thus, it is not important for the allocation to be exactly optimal.

In Figure 1, when $n = 500$, the optimal allocation's SE, 0.230, is about 15% lower than the SE that could be achieved without any baseline interviews, 0.272. Therefore, the minimum detectable effect (MDE; Bloom, 1995) is about 15% smaller under the optimal allocation than it would be without any baseline interviews. (When $n = 100$, the corresponding reduction is about 13%.) For example, for a two-sided test at the 10% significance level, the MDE with 80% power is $2.49 \times 0.230 = 0.573$ under the optimal allocation, while it is $2.49 \times 0.272 = 0.677$ without any baseline interviews. (The unit for these MDEs is a point on the 5-point scale of 1 = "*much worse*" to 5 = "*much better*" for the economic optimism question.)

Figure 2 plots the analogous calculations for the protest inclination outcome variable. (The SEs are much smaller because the estimate of the between-cluster variance $\sigma^2_{y(\text{between})}$ is only 0.0325 for this variable.) When $n = 500$, the optimal allocation ($\pi_{\text{opt}} = 0.386$) achieves an SE of 0.066, which is about 20% lower than the SE that could be achieved without any baseline interviews (0.082). However, as discussed in note 11, when n is reduced to 100, the usefulness of the baseline covariate data declines substantially because the ICC for inclination to protest is not very large. Now the optimal allocation ($\pi_{\text{opt}} = 0.288$) achieves an SE of 0.084, which is only about 6% lower than the SE that could be achieved without any baseline interviews (0.089). Again, it is not important for the allocation to be exactly optimal. When $n = 500$, all baseline allocations between 20% and 50% achieve approximately the same precision. When $n = 100$, a 50% baseline allocation results in an SE of 0.086, which is just slightly higher than the optimal SE (0.084).

Imbalanced Designs

Many experiments and quasi-experiments use imbalanced designs with unequal-sized treatment and control groups. For example, if the intervention is very costly, the researchers may decide to assign $J_t < J/2$ clusters to treatment and $J_c = J - J_t$ clusters to control. In this section, we assume that

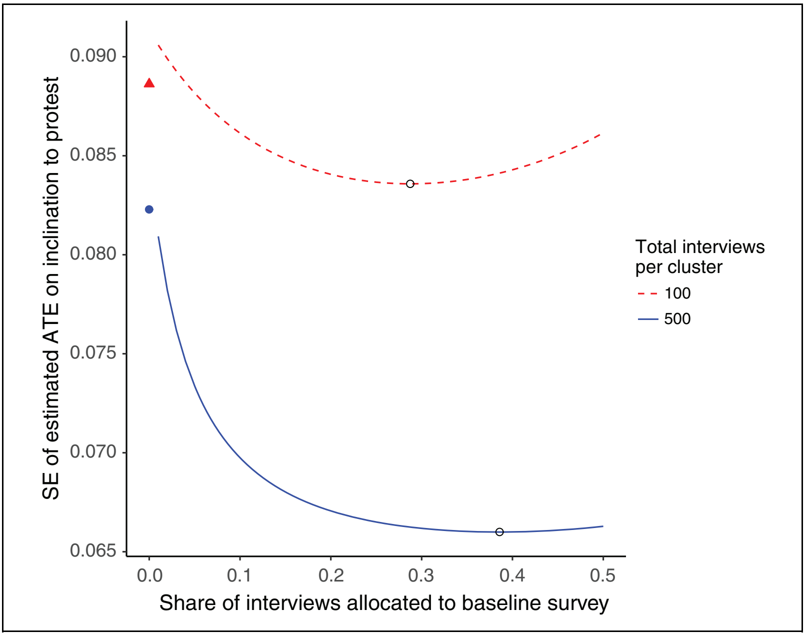

Figure 2. Survey allocation and precision when the outcome variable is inclination to protest. Near the top left corner, the filled triangle (for $n = 100$ interviews per cluster) and circle (for $n = 500$) show the standard error (SE) of the unadjusted estimate of average treatment effect when all interviews are allocated to the endline survey. The curves plot the SE of the regression-adjusted estimate against the share of interviews allocated to the baseline survey. The open circle on each curve marks the optimal baseline share. See text for details.

J, J_t, and the total number of survey interviews S have already been chosen, but we need to decide how to allocate the S interviews between the baseline and endline surveys and between the treatment and control group clusters. (As shown below, it turns out to be desirable to allocate more interviews per cluster to the group that has fewer clusters.) We assume here that the baseline survey, if any, will be administered after clusters are assigned but before treatment begins. Thus, for both the baseline survey and the endline survey, we have the option of allocating different numbers of interviews per cluster to the treatment and control groups.

We also assume that if a baseline survey is conducted, we will estimate the ATE using the coefficient on T_j in an ordinary least squares regression of $\widehat{Y}_j$ on T_j, $\widehat{X}_j$, and the interaction $T_j \cdot (\widehat{X}_j - \frac{1}{J}\sum_{k=1}^{J}\widehat{X}_k)$.

Including the interaction can improve asymptotic precision in imbalanced designs (Lin, 2013; Yang & Tsiatis, 2001). In our context, the interaction term allows the regression model to take into account the possibility that the correlation between $\widehat{X}_j$ and $\widehat{Y}_j$ is stronger or weaker in the treatment group than the control group. For example, if we allocate more baseline and endline interviews per cluster to the treatment group than to the control group, then the cluster-level sample means $\widehat{X}_j$ and $\widehat{Y}_j$ will be noisier estimates of the cluster-level population means X_j and Y_j in the control group than in the treatment group. We would therefore expect the correlation between $\widehat{X}_j$ and $\widehat{Y}_j$ to be stronger in the treatment group than in the control group.

While it appears to be difficult to solve for an exact optimum, numerical calculations (such as those in the example below) suggest that the following allocation performs well in many scenarios:

- Allocate half the interviews to the treatment group and half to the control group.[12] The number of interviews *per cluster* will then differ between the treatment and control groups: There will be $n_t = S/(2J_t)$ interviews per cluster in the treatment group and $n_c = S/(2J_c)$ in the control group. For example, if the budget allows $S = 20,000$ interviews, and there are 30 clusters with 10 assigned to treatment and 20 to control, then allocate 10,000 interviews (1,000 per cluster) to the treatment group and 10,000 (500 per cluster) to the control group.
- Let $n_m = \max(n_t, n_c)$. If $|\rho| \leq K/n_m$, allocate all interviews to the endline survey. If $|\rho| > K/n_m$, allocate a proportion of interviews

$$\pi = 1 - \frac{1 + K/n_m}{1 + |\rho|} \tag{11}$$

to the baseline survey. (Although π could be allowed to differ between the treatment and control groups, in many scenarios, there is little gain from such fine-tuning. The suggested baseline allocation here mimics the one we derived for the balanced design in Equation 9 and uses n_m, the number of interviews per cluster in the group that has fewer clusters.)

Example: A Digital Advertising Experiment

To illustrate the ideas discussed above, we consider an application to digital political advertising drawn from Turitto, Green, Stobie, and Tranter (2014). Ten of 30 noncontiguous midsized cities in Texas were randomly assigned to the treatment, a 7-day digital advertising campaign on behalf of David Dewhurst, the incumbent candidate for lieutenant governor in the 2014 Republican primary. Using a repeated cross-section design, a baseline survey of Republican voters was conducted during January 3–6 (just before the launch of the treatment) and an endline survey was conducted during January 14–17 (just after the treatment ended). These automated phone surveys asked respondents, "Thinking about the race for Texas Lieutenant Governor for a moment, if the primary election were held today, which of the following candidates would you vote for?" and presented a list of candidates in random order. The goal of the study was to estimate the effect of the treatment on the proportion of respondents who indicated that they would vote for Dewhurst.

The baseline survey was designed to obtain approximately 100 interviews in each treatment group city and 50 interviews in each control group city, while the endline survey was designed to obtain approximately 300 interviews in each treatment group city and 150 interviews in each control group city. Thus, out of a total of approximately 8,000 interviews, half were allocated to the treatment group and half to the control group (as suggested above), with 25% allocated to the baseline survey and 75% to the endline survey.

We can explore in hindsight how the precision of the estimated treatment effect would vary with alternative allocations of the survey interviews.[13] The budgeted total number of interviews is $S = 8,000$, with $J_t = 10$ cities in the treatment group and $J_c = 20$ cities in the control group. Our outcome variable is defined as 1 if the respondent indicated support for Dewhurst and 0 otherwise. Using the endline survey data and the same methods as in the Afrobarometer example, we estimate $\sigma^2_{y(\mathrm{between})}$ as 0.00575 and ICC_y as 0.0291. The corresponding ICC estimate from the baseline survey is 0.0210. The observed correlation between the baseline and endline city-level means of the outcome variable is 0.609, and the harmonic means of the city-level sample sizes are 56.9 for the baseline survey and 175.7 for the endline survey. Applying Equation 10, we estimate the covariate–outcome correlation ρ as 0.895 (which suggests that city-level support for Dewhurst was fairly stable from early to mid-January).

Next, we calculate the suggested proportion of baseline interviews π from Equation 11. The suggested number of interviews per city is $n_t = S/(2J_t) = 400$ for treatment group cities and $n_c = S/(2J_c) = 200$ for control group cities, so the parameter n_m in (11) equals 400. The boundary condition $|\rho| > K/n_m$ is easily satisfied with the above estimates of ρ and the ICC. Equation 11 yields $\pi = 0.43$.

Figure 3 explores how alternative allocations of the survey interviews would affect the SE of the estimated treatment effect.[14] Each curve shows how the SE varies with the share of interviews allocated to the treatment group, holding the share allocated to the baseline survey (which is assumed to be the same across the treatment and control groups) constant at zero, 25% (the actual baseline share), 43% (the baseline share suggested above), or 50%. Comparisons within each curve show that the SE is minimized when half the interviews are allocated to the treatment group and half to the control group. Comparisons across the bottom three curves show that precision is only slightly better with the suggested 43% baseline share (yielding at best an SE of 2.28 percentage points, which implies an MDE of 5.68 percentage points) than a 25% or 50% baseline share (yielding an SE of 2.35 or 2.30 percentage points at best, which implies an MDE of 5.85 or 5.73 percentage points). Thus, the actual 25% baseline share appears to have been a reasonable choice. Finally, the topmost (dotted) curve shows that precision would be noticeably worse if all interviews were allocated to the endline survey (the SE for the unadjusted estimate is at best 3.1 percentage points, implying an MDE of 7.7 percentage points).

Discussion

When the outcomes of interest are relatively stable over time, a study design with repeated cross-section surveys can be an effective strategy for efficient estimation of ATEs. Our analysis is intended to sketch some of the key issues involved in cost-efficient allocation of survey interviews and to invite more complex formalizations of the allocation problem. For simplicity, we omitted a number of complications that researchers may want to consider in applications, such as multiple baseline or follow-up survey waves, fixed costs associated with each survey wave, asymmetric costs of interviews in treatment and control areas, use of multiple baseline covariates in regression adjustment, and motivations for conducting a baseline survey other than improving the precision of estimated ATEs. Also, we assumed that the estimand is an ATE that weights each cluster equally, but researchers may prefer to weight the clusters according to population size or other considerations. Furthermore, we took the numbers of clusters assigned to treatment and control as given, while a more sophisticated analysis would simultaneously optimize the allocation of clusters to treatment arms and the allocation of survey interviews, given information about treatment costs, survey costs, and the overall budget. Researchers may wish to use our framework as a starting point for more complicated analyses that consider such issues.

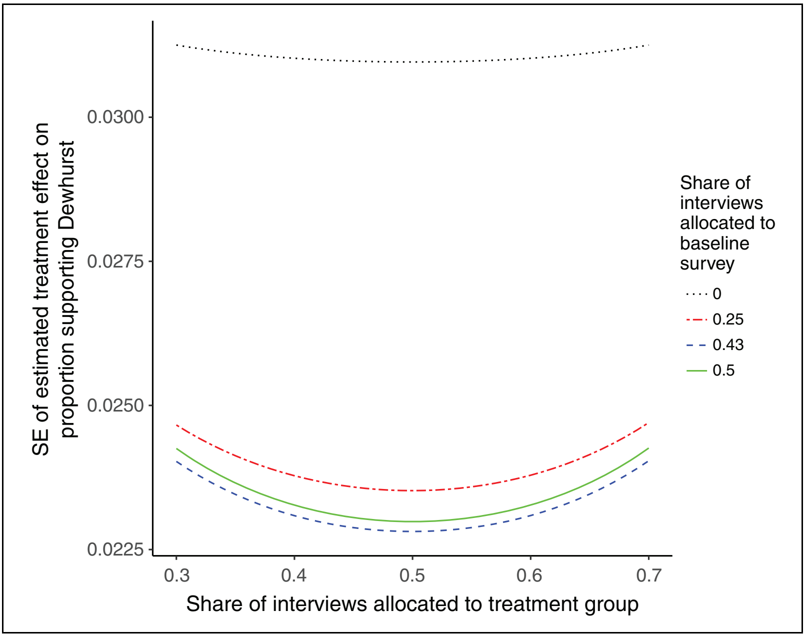

Figure 3. Survey allocation and precision in the digital advertising example. The top (dotted) curve plots the SE of the unadjusted treatment effect estimate against the treatment group's share of interviews when all interviews are allocated to the endline survey. The other three curves plot the SE of the regression-adjusted estimate against the treatment group's share of interviews, holding the baseline survey's share constant at 25% (the actual allocation), 43% (the suggested allocation), or 50%. See text for details.

Because we omitted such complications, the formulas we have given for optimal allocation will not necessarily be optimal in practice, but the analysis may be of heuristic value. In a cluster-randomized experiment with repeated cross-section surveys, the optimal share of interviews to allocate to the baseline survey is less than one-half and tends to increase with the cluster-level correlation between baseline and endline measures of the outcome variable, the ICC, and the total number of interviews per cluster. In many scenarios, a wide range of baseline allocations yields approximately the same statistical precision. This suggests that researchers have quite a bit of latitude to accommodate other design considerations, such as fielding a baseline survey in order to train enumerators or pretest a survey instrument.

Appendix

Balanced Designs

Derivation of formula (7). Let $\widehat{Y}_j(0)$ denote the value that $\widehat{Y}_j$ would take if cluster j were assigned to control:

$$\widehat{Y}_j(0) = \frac{1}{n_{\text{post}}} \sum_{i \in E_j} y_{ij}(0). \tag{A1}$$

Similarly, let $\widehat{Y}_j(1)$ denote the value that $\widehat{Y}_j$ would take if cluster j were assigned to treatment:

$$\widehat{Y}_j(1) = \frac{1}{n_{\text{post}}} \sum_{i \in E_j} y_{ij}(1) = \widehat{Y}_j(0) + \tau. \tag{A2}$$

Recall that the J clusters are assumed to be randomly drawn from a much larger superpopulation. If the sampling fraction (J divided by the number of clusters in the superpopulation) is close to zero, then, applying a basic result on the variance of the treatment–control difference in mean outcomes (Imbens & Rubin, 2015, p. 101, eq. 6.17), we get

$$\text{Var}(\widehat{\text{ATE}}_{\text{unadj}}) \approx \frac{\text{Var}(\widehat{Y}_j(0))}{J/2} + \frac{\text{Var}(\widehat{Y}_j(1))}{J/2} = \frac{4}{J}\text{Var}(\widehat{Y}_j(0)). \tag{A3}$$

For each cluster j, let

$$\sigma^2_{y(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} (y_{ij}(0) - Y_j(0))^2. \tag{A4}$$

By the law of total variance,

$$\text{Var}(\widehat{Y}_j(0)) = \text{Var}\left[E(\widehat{Y}_j(0) | Y_j(0), \sigma^2_{y(j)})\right] + E\left[\text{Var}(\widehat{Y}_j(0) | Y_j(0), \sigma^2_{y(j)})\right] \tag{A5}$$

$$= \text{Var}(Y_j(0)) + E(\sigma^2_{y(j)}/n_{\text{post}}) \tag{A6}$$

$$= \sigma^2_{y(\text{between})} + \sigma^2_{y(\text{within})}/n_{\text{post}} \tag{A7}$$

$$= \sigma^2_{y(\text{between})}(1 + K/n_{\text{post}}), \tag{A8}$$

so

$$\text{Var}(\widehat{\text{ATE}}_{\text{unadj}}) \approx \frac{4}{J}\sigma^2_{y(\text{between})}(1 + K/n_{\text{post}}). \tag{A9}$$

Derivation of formula (8). Equation 7 of Yang and Tsiatis (2001) implies that as J goes to infinity, the asymptotic variance of $\widehat{\text{ATE}}_{\text{adj}}$ is

$$\begin{aligned}\frac{1}{J}\Bigg[&\frac{1}{1-0.5}\text{Var}(\widehat{Y}_j|T_j=0)+\frac{1}{0.5}\text{Var}(\widehat{Y}_j|T_j=1)\\&+\frac{1}{\text{Var}(\widehat{X}_j)\cdot 0.5\cdot(1-0.5)}\left\{(1-0.5)\cdot\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=0)+0.5\cdot\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=1)\right\}\\&\times\left\{(1-3\cdot 0.5)\cdot\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=0)+(3\cdot 0.5-2)\cdot\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=1)\right\}\Bigg].\end{aligned} \tag{A10}$$

(An exact variance formula would require stronger assumptions, such as correct specification of the regression model.)

Since T_j is randomly assigned and we assume a homogeneous treatment effect, the conditional variances and covariances in formula (A10) can be reexpressed as follows:

$$\text{Var}(\widehat{Y}_j|T_j=0)=\text{Var}(\widehat{Y}_j(0)|T_j=0)=\text{Var}(\widehat{Y}_j(0)), \tag{A11}$$

$$\text{Var}(\widehat{Y}_j|T_j=1)=\text{Var}(\widehat{Y}_j(1)|T_j=1)=\text{Var}(\widehat{Y}_j(1))=\text{Var}(\widehat{Y}_j(0)), \tag{A12}$$

$$\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=0)=\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0)|T_j=0)=\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0)), \tag{A13}$$

$$\begin{aligned}&\text{Cov}(\widehat{X}_j,\widehat{Y}_j|T_j=1)=\text{Cov}(\widehat{X}_j,\widehat{Y}_j(1)|T_j=1)=\text{Cov}(\widehat{X}_j,\widehat{Y}_j(1))\\&=\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0)).\end{aligned} \tag{A14}$$

Thus, formula (A10) simplifies to

$$\text{Avar}(\widehat{\text{ATE}}_{\text{adj}})=\frac{1}{J}\left[4\text{Var}(\widehat{Y}_j(0))-4\frac{\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0))^2}{\text{Var}(\widehat{X}_j)}\right]=\frac{4}{J}\text{Var}(\widehat{Y}_j(0))(1-\rho_*^2), \tag{A15}$$

where

$$\rho_*^2=\frac{\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0))^2}{\text{Var}(\widehat{X}_j)\text{Var}(\widehat{Y}_j(0))}. \tag{A16}$$

Since $\widehat{X}_j$ and $\widehat{Y}_j(0)$ are just X_j and $Y_j(0)$ plus independent random sampling errors, $\text{Cov}(\widehat{X}_j,\widehat{Y}_j(0))=\text{Cov}(X_j,Y_j(0))$. Thus,

$$\rho_*^2=\frac{\text{Cov}(X_j,Y_j(0))^2}{\text{Var}(\widehat{X}_j)\text{Var}(\widehat{Y}_j(0))}=\rho^2\frac{\sigma^2_{x(\text{between})}}{\text{Var}(\widehat{X}_j)}\frac{\sigma^2_{y(\text{between})}}{\text{Var}(\widehat{Y}_j(0))}. \tag{A17}$$

Using Equation (A8) and the analogous result for $\mathrm{Var}(\widehat{X}_j)$, we get

$$\rho_*^2 = \rho^2 \left(1 + \frac{K}{n_{\text{pre}}}\right)^{-1} \left(1 + \frac{K}{n_{\text{post}}}\right)^{-1}. \tag{A18}$$

Plug Equations (A8) and (A18) into Equation (A15) to get

$$\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{adj}}) = \frac{4}{J}\sigma^2_{y(\text{between})}\left(1 + \frac{K}{n_{\text{post}}}\right)\left[1 - \rho^2\left(1 + \frac{K}{n_{\text{pre}}}\right)^{-1}\left(1 + \frac{K}{n_{\text{post}}}\right)^{-1}\right]. \tag{A19}$$

Optimal allocation. Reexpressing Equation (A19) in terms of $n = n_{\text{pre}} + n_{\text{post}}$ and $\pi = n_{\text{pre}}/n$, we get

$$\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{adj}}) = \frac{4}{J}\sigma^2_{y(\text{between})}\left(1 + \frac{K}{(1-\pi)n}\right)\left[1 - \rho^2\left(1 + \frac{K}{\pi n}\right)^{-1}\left(1 + \frac{K}{(1-\pi)n}\right)^{-1}\right]. \tag{A20}$$

Thus, choosing $\pi \in (0, 1)$ to minimize $\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{adj}})$ is equivalent to minimizing

$$g(\pi) = \frac{K}{(1-\pi)n} - \rho^2(1 + \frac{K}{\pi n})^{-1}, \tag{A21}$$

which has first and second derivatives

$$g'(\pi) = \frac{K}{n}\left[\frac{1}{(1-\pi)^2} - \left(\frac{\rho n}{\pi n + K}\right)^2\right] \tag{A22}$$

and

$$g''(\pi) = \frac{2K}{n}\left[\frac{1}{(1-\pi)^3} + \frac{\rho^2 n^3}{(\pi n + K)^3}\right] > 0. \tag{A23}$$

If $|\rho| > K/n$, then the value $\pi_{\text{opt}} = 1 - (1 + K/n)/(1 + |\rho|)$ satisfies both $0 < \pi_{\text{opt}} < 1$ and $g'(\pi_{\text{opt}}) = 0$, so $\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{adj}})$ is minimized at $\pi = \pi_{\text{opt}}$. It can be shown that the resulting value for $\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{adj}})$ (plug $\pi = \pi_{\text{opt}}$ into Equation [A20]) is lower than the lowest possible value for $\mathrm{Var}(\widehat{\mathrm{ATE}}_{\mathrm{unadj}})$ (plug $n_{\text{post}} = n$ into Equation [A9]). Thus, for large enough J, the optimal allocation is $\pi = \pi_{\text{opt}}$.

If $|\rho| \leq K/n$, then $g(\pi)$ is an increasing function for $0 < \pi < 1$, and the optimal allocation is $\pi = 0$ (allocate all interviews to the endline survey and use $\widehat{\text{ATE}}_{\text{unadj}}$).

Estimation of ρ *for Illustrative Examples*

Equation (A18) gives the relationship between ρ_* (the correlation between $\widehat{X}_j$ and $\widehat{Y}_j(0)$) and ρ (the correlation between X_j and $Y_j(0)$) when the sample sizes n_{pre} and n_{post} are constant across clusters. When the sample sizes vary by cluster (as in the data sets for our illustrative examples), the law of total variance yields

$$\text{Var}(\widehat{Y}_j(0)) = \text{Var}[E(\widehat{Y}_j(0)|Y_j(0), \sigma^2_{y(j)})] + E[\text{Var}(\widehat{Y}_j(0)|Y_j(0), \sigma^2_{y(j)})] \quad \text{(A24)}$$

$$= \text{Var}(Y_j(0)) + E(\sigma^2_{y(j)}/n_{j,\text{post}}) \quad \text{(A25)}$$

and similarly

$$\text{Var}(\widehat{X}_j) = \text{Var}(X_j) + E(\sigma^2_{x(j)}/n_{j,\text{pre}}), \quad \text{(A26)}$$

where $n_{j,\text{pre}}$ and $n_{j,\text{post}}$ are the sample sizes for cluster j. One possible approach is to randomly delete observations in such a way as to make $n_{j,\text{pre}}$ and $n_{j,\text{post}}$ constant across clusters, then estimate ρ_* and K, and finally use Equation (A18) to estimate ρ. Alternatively, one could estimate $\sigma^2_{x(j)}/n_{j,\text{pre}}$ and $\sigma^2_{y(j)}/n_{j,\text{post}}$ separately for each cluster and then take means across clusters to estimate the expectations. We use a simpler calculation that assumes independence between $\sigma^2_{x(j)}$ and $n_{j,\text{pre}}$ and between $\sigma^2_{y(j)}$ and $n_{j,\text{post}}$. In that case,

$$\begin{aligned}\text{Var}(\widehat{Y}_j(0)) &= \text{Var}(Y_j(0)) + E(\sigma^2_{y(j)}) \cdot E(n^{-1}_{j,\text{post}}) && \text{(A27)}\\ &= \sigma^2_{y(\text{between})} + \sigma^2_{y(\text{within})} \cdot E(n^{-1}_{j,\text{post}}) && \text{(A28)}\\ &= \sigma^2_{y(\text{between})}[1 + K \cdot E(n^{-1}_{j,\text{post}})] && \text{(A29)}\end{aligned}$$

and similarly $\text{Var}(\widehat{X}_j) = \sigma^2_{x(\text{between})}[1 + K \cdot E(n^{-1}_{j,\text{pre}})]$, so Equation (A17) implies

$$\rho^2 = \rho^2_*[1 + K \cdot E(n^{-1}_{j,\text{pre}})][1 + K \cdot E(n^{-1}_{j,\text{post}})]. \quad \text{(A30)}$$

We use the sample correlation ρ_{obs} to estimate ρ_*; K_4 times the sample mean of $n^{-1}_{j,\text{pre}}$ to estimate $K \cdot E(n^{-1}_{j,\text{pre}})$; and K_5 times the sample mean of $n^{-1}_{j,\text{post}}$ to estimate $K \cdot E(n^{-1}_{j,\text{post}})$.

Imbalanced Designs

Let $\widehat{\mathrm{ATE}}_{\mathrm{interact}}$ denote the estimated coefficient on T_j in the ordinary least squares regression of $\widehat{Y}_j$ on T_j, $\widehat{X}_j$, and $T_j \cdot (\widehat{X}_j - \frac{1}{J}\sum_{k=1}^{J}\widehat{X}_k)$. Using arguments similar to those in Pitkin et al. (2017, proof of Lemma 3.4), it can be shown that as J goes to infinity, the asymptotic variance of $\widehat{\mathrm{ATE}}_{\mathrm{interact}}$ is

$$\begin{aligned}
&\mathrm{Avar}(\widehat{\mathrm{ATE}}_{\mathrm{interact}}) \\
&= \sigma^2_{y(\mathrm{between})}\left\{\frac{1}{J_t}\left(1+\frac{K}{(1-\pi)n_t}\right)\left[1-\rho^2\left(1+\frac{K}{\pi n_t}\right)^{-1}\left(1+\frac{K}{(1-\pi)n_t}\right)^{-1}\right]\right. \\
&\quad +\frac{1}{J_c}\left(1+\frac{K}{(1-\pi)n_c}\right)\left[1-\rho^2\left(1+\frac{K}{\pi n_c}\right)^{-1}\left(1+\frac{K}{(1-\pi)n_c}\right)^{-1}\right] \\
&\quad \left.+\frac{\rho^2}{J^2}\left[\left(1+\frac{K}{\pi n_t}\right)^{-1}-\left(1+\frac{K}{\pi n_c}\right)^{-1}\right]^2\times\left[J_t\left(1+\frac{K}{\pi n_t}\right)+J_c\left(1+\frac{K}{\pi n_c}\right)\right]\right\}.
\end{aligned} \tag{A31}$$

In Figure 3, the approximate SE of $\widehat{\mathrm{ATE}}_{\mathrm{interact}}$ is calculated by multiplying the above formula by the degrees-of-freedom correction $(J-3)/(J-5)$ and then taking the square root.

For the variance of $\widehat{\mathrm{ATE}}_{\mathrm{unadj}}$ when all interviews are allocated to the endline survey, the derivation is similar to that in Equations (A1)–(A9) but slightly more complicated. If cluster j is assigned to treatment, the endline survey collects data from a random sample $E_j(1)$ of n_t individuals; if the cluster is assigned to control, the survey collects data from a random sample $E_j(0)$ of n_c individuals. In place of Equations (A1) and (A2), we have

$$\widehat{Y}_j(0) = \frac{1}{n_c}\sum_{i\in E_j(0)} y_{ij}(0) \tag{A32}$$

and

$$\widehat{Y}_j(1) = \frac{1}{n_t}\sum_{i\in E_j(1)} y_{ij}(1) = \frac{1}{n_t}\sum_{i\in E_j(1)} y_{ij}(0) + \tau. \tag{A33}$$

The basic result on the variance of the treatment–control difference in mean outcomes (Imbens & Rubin, 2015, p. 101, eq. 6.17) yields

$$\mathrm{Var}(\widehat{\mathrm{ATE}}_{\mathrm{unadj}}) \approx \frac{\mathrm{Var}(\widehat{Y}_j(0))}{J_c} + \frac{\mathrm{Var}(\widehat{Y}_j(1))}{J_t}. \tag{A34}$$

By the law of total variance,

$$\mathrm{Var}(\widehat{Y}_j(0)) = \mathrm{Var}[E(\widehat{Y}_j(0)|Y_j(0), \sigma^2_{y(j)})] + E[\mathrm{Var}(\widehat{Y}_j(0)|Y_j(0), \sigma^2_{y(j)})] \quad \text{(A35)}$$

$$= \mathrm{Var}(Y_j(0)) + E(\sigma^2_{y(j)}/n_c) \quad \text{(A36)}$$

$$= \sigma^2_{y(\text{between})} + \sigma^2_{y(\text{within})}/n_c \quad \text{(A37)}$$

$$= \sigma^2_{y(\text{between})}(1 + K/n_c) \quad \text{(A38)}$$

and similarly $\mathrm{Var}(\widehat{Y}_j(1)) = \sigma^2_{y(\text{between})}(1 + K/n_t)$. Thus,

$$\mathrm{Var}(\widehat{\mathrm{ATE}}_{\text{unadj}}) \approx \sigma^2_{y(\text{between})}\left[\frac{1 + K/n_c}{J_c} + \frac{1 + K/n_t}{J_t}\right]. \quad \text{(A39)}$$

When there is no baseline survey, the optimal allocation of interviews between the treatment and control groups chooses n_t and n_c to minimize $\mathrm{Var}(\widehat{\mathrm{ATE}}_{\text{unadj}})$ subject to the constraint that $J_t n_t + J_c n_c = S$. Rewriting Equation (A39) as

$$\mathrm{Var}(\widehat{\mathrm{ATE}}_{\text{unadj}}) \approx \sigma^2_{y(\text{between})}\left[\frac{1}{J_c} + \frac{1}{J_t} + K\left(\frac{1}{J_c n_c} + \frac{1}{J_t n_t}\right)\right] \quad \text{(A40)}$$

the only part that depends on n_t and n_c is

$$\frac{1}{J_c n_c} + \frac{1}{J_t n_t}, \quad \text{(A41)}$$

which is minimized when $J_c n_c = J_t n_t = S/2$. Thus, it is optimal to allocate half the interviews to the treatment group and half to the control group.

Acknowledgments

The authors are grateful to Jacob Klerman (the editor), Peter Aronow, and three anonymous reviewers for very helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For automated phone surveys, the costs of repeated cross-sectional interviewing are essentially identical between baseline and endline. In-person survey costs

are also likely to be similar across waves (perhaps with slightly lower unit costs in the endline if the survey team has become more experienced and efficient).

2. Our setup omits some considerations that may be important in practice. In many applications, the relevant costs will also depend on other factors, such as the number of clusters, in total and in the treatment group; the number of interviews allocated to each cluster and each survey; and fixed costs of each survey wave (which may be considerable). Thus, the budget constraint will involve a cost function that depends on more than one variable, and if the fixed costs are high enough, it may be optimal to dispense with the baseline wave entirely. On the other hand, there may be other reasons to conduct a baseline survey besides precision improvement (piloting questions, describing the population, etc.), so the objective function may depend on other factors besides the precision of the estimated average treatment effect. We return to these considerations in the concluding section.
3. In large samples, the gains from blocking are similar to those from poststratification, which is in turn a form of linear regression adjustment (Miratrix, Sekhon, & Yu, 2013).
4. In randomized experiments, linear regression adjustment with robust standard errors can be used to construct asymptotically valid confidence intervals for average treatment effects even when the regression model is misspecified (Lin, 2013). Judkins and Porter (2016) found in simulations that ordinary least squares regression adjustment gave valid inferences even with a binary outcome and very small sample sizes. When the goal is to estimate the ATE, nonlinear models such as logit and probit require the additional complexity of average marginal effect calculations (Angrist & Pischke, 2009, pp. 103–107; Freedman, 2008; Williams, 2012), and estimates based on the probit maximum likelihood estimator are not misspecification-robust (Firth & Bennett, 1998; Freedman, 2008).
5. When the baseline and endline samples are drawn independently, there is a chance that some individuals will be selected for both surveys, but the overlap is likely to be small if the sample sizes are small relative to the population size. A reviewer helpfully pointed out that there is a literature on nonindependent sampling algorithms that minimize or maximize overlap (Ernst, 1998). It is possible that an overlap-minimizing algorithm could be useful (to minimize the priming effects we mentioned in the introduction) in some applications with small population sizes.
6. Raudenbush (1997, pp. 181–182) shows that this “aggregated” regression gives the same treatment effect estimate as a multilevel model that allows the between-cluster relationship between the covariate and the outcome to differ from the within-cluster relationship.

7. To understand the role of the intracluster correlation coefficient (ICC), note that random treatment–control differences in the sample means of the covariate and potential outcomes are due to both (1) the random or quasi-random assignment of whole clusters to treatment and (2) the random sampling of individuals within each cluster for the baseline and endline surveys. In the case where the ICC is near 0, the between-cluster variances of the covariate and potential outcomes are very small relative to the within-cluster variances, so the treatment–control differences are largely driven by the random sampling of individuals for the surveys. Since the baseline and endline survey samples are independent, the baseline difference is of little value for predicting the endline difference in this case. But in the case where the ICC is near 1, the between-cluster variances are very large relative to the within-cluster variances, so the treatment–control differences are largely driven by the random or quasi-random assignment of clusters, and if $\rho \neq 0$, then the baseline difference in sample mean covariate values does help predict the endline difference in sample mean outcomes.
8. The countries are Benin, Botswana, Burkina Faso, Cape Verde, Ghana, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, and Zimbabwe.
9. The other response categories were 2 = "*worse*," 3 = "*same*," and 4 = "*better*." We omitted don't-knows, refusals, and missing values from the analysis.
10. The other response categories were 1 = "*no, but would do if had the chance*," 2 = "*yes, once or twice*," and 3 = "*yes, several times*." Again, we omitted don't-knows, refusals, and missing values.
11. Thus, the optimal baseline share is more sensitive to changes in n when we use the estimates of ρ and K for inclination to protest. The reason is that this variable is estimated to have a larger true covariate–outcome correlation ρ and a smaller ICC than the economic optimism variable. As explained in our discussion of formula (9), increases in n and the ICC improve the signal-to-noise ratio in the sample means of the covariate and outcome at the cluster level. When $n = 500$, the measurement noise in the sample means is less of an issue, so the optimal baseline share π_{opt} is greater for the outcome variable with the larger true ρ (inclination to protest). But when n is reduced to 100, the usefulness of the baseline data declines more substantially for the inclination-to-protest variable because it has a smaller ICC, so now π_{opt} is smaller for inclination to protest than for economic optimism.
12. The Appendix shows that this half-half allocation minimizes the variance of the unadjusted treatment effect estimate (Equation A39). Although we have not found a solution to the problem of minimizing the asymptotic variance of the regression-adjusted estimate (equation [A31] in the Appendix), the half–half

allocation minimized this expression not only in Figure 3, but also when we changed the ICC to values ranging from 0.001 to 0.9, when we reduced ρ from 0.895 to 0.1, and when we reduced the total number of interviews *S* from 8,000 to 800.

13. For simplicity, this illustrative example omits some details of the original study. In addition to the baseline and endline surveys, the researchers had access to the Dewhurst campaign's polls from November and December, which were used to group the cities into 10 blocks of 3 cities each. Within each block, one city was randomly assigned to the treatment. As a result, the study achieved greater precision than shown in our example. The reported SE of the estimated treatment effect was 2.1 percentage points.
14. Equations (A31) and (A39) in the Appendix give the formulas we used to calculate the SEs of the regression-adjusted and unadjusted estimates.

References

Afrobarometer. (2009). *Merged round 4 data (20 countries) (2008)* [Data file]. Retrieved from http://www.afrobarometer.org

Afrobarometer. (2015). *Merged round 5 data (34 countries) (2011–2013) (last update: July 2015)* [Data file]. Retrieved from http://www.afrobarometer.org

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*, 547–556.

Bloom, H. S., & Riccio, J. A. (2005). Using place-based random assignment and comparative interrupted time-series analysis to evaluate the Jobs-Plus employment program for public housing residents. *Annals of the American Academy of Political and Social Science*, *599*, 19–51.

Boruch, R. (2005). Better evaluation for evidence-based policy: Place randomized trials in education, criminology, welfare, and health. *Annals of the American Academy of Political and Social Science*, *599*, 6–18.

Cheadle, A., Psaty, B. M., Diehr, P., Koepsell, T., Wagner, E., Curry, S., & Kristal, A. (1995). Evaluating community-based nutrition programs: Comparing grocery store and individual-level survey measures of program impact. *Preventive Medicine*, *24*, 71–79.

Cox, D. R., & McCullagh, P. (1982). Some aspects of analysis of covariance. *Biometrics*, *38*, 541–561.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, *54*, 67–82.

Dorn, D., Fischer, J. A. V., Kirchgässner, G., & Sousa-Poza, A. (2007). Is it culture or democracy? The impact of democracy and culture on happiness. *Social Indicators Research*, *82*, 505–526.

Ernst, L. R. (1998). Maximizing and minimizing overlap when selecting a large number of units per stratum simultaneously for two designs. *Journal of Official Statistics*, *14*, 297–314.

Farquhar, J. W., Fortmann, S. P., Maccoby, N., Haskell, W. L., Williams, P. T., & Flora, J. A., . . . Hulley, S. B. (1985). The Stanford five-city project: Design and methods. *American Journal of Epidemiology*, *122*, 323–334.

Firth, D., & Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*, 3–21.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London: Series A (Containing Papers of a Mathematical or Physical Character)*, *222*, 309–368.

Flay, B. R., & Collins, L. M. (2005). Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *Annals of the American Academy of Political and Social Science*, *599*, 115–146.

Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science*, *23*, 237–249.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, *15*, 1069–1092.

Gerber, A. S., Gimpel, J. G., Green, D. P., & Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *American Political Science Review*, *105*, 135–150.

Green, D. P., Wilke, A., Cooper, J., & Baltes, S. (2016). Can media shape social norms? A randomized experiment assessing portrayals of domestic violence, abortion, and teacher absenteeism in rural Uganda. Paper presented at the EGAP Conference, New Haven, CT, October 14–15, Yale University.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, *81*, 945–970.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York, NY: Cambridge University Press.

Judkins, D. R., & Porter, K. E. (2016). Robustness of ordinary least squares in randomized clinical trials. *Statistics in Medicine*, *35*, 1763–1773.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, *7*, 295–318.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*, 210–221.

Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*, 369–396.

Murray, D. M., Hannan, P. J., Jacobs, D. R., McGovern, P. J., Schmid, L., Baker, W. L., & Gray, C. (1994). Assessing intervention effects in the Minnesota heart health program. *American Journal of Epidemiology*, *139*, 91–103.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. In D. M. Dabrowska & T. P. Speed (Eds. & Trans.), (with discussion) in *Statistical Science*, *5*, 463–480.

Parker, S. W., & Teruel, G. M. (2005). Randomization and social program evaluation: The case of Progresa. *Annals of the American Academy of Political and Social Science*, *599*, 199–219.

Pitkin, E., Berk, R., Brown, L., Buja, A., George, E., Zhang, K., & Zhao, L. (2017). An asymptotically powerful test for the average treatment effect. Retrieved from http://www-stat.wharton.upenn.edu/∼lbrown/

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*, 117–128.

Ridout, M. S., Demétrio, C. G. B., & Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, *55*, 137–148.

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

Smith, H. L., Ping, T., Merli, M. G., & Hereward, M. (1997). Implementation of a demographic and contraceptive surveillance system in four counties in north China. *Population Research and Policy Review*, *16*, 289–314.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*, 137–150.

Ter Kuile, F. O., Terlouw, D. J., Phillips-Howard, P. A., Hawley, W. A., Friedman, J. F., & Kolczak, M. S., … Nahlen, B. L. (2003). Impact of permethrin-treated bed nets on malaria and all-cause morbidity in young children in an area of intense perennial malaria transmission in western Kenya: Cross-sectional survey. *American Journal of Tropical Medicine and Hygiene*, *68*, 100–107.

Turitto, C., Green, D. P., Stobie, B., & Tranter, S. (2014, August 30). Testing the persuasive effects of digital media: A cluster randomized field experiment. Paper presented at the Annual Meeting of the American Political Science Association. Washington, DC.

Welsch, H. (2007). Macroeconomics and life satisfaction: Revisiting the “misery index.” *Journal of Applied Economics*, *10*, 237–251.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal*, *12*, 308–331.

Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *American Statistician*, *55*, 314–321.

Author Biographies

Donald P. Green is J. W. Burgess Professor of Political Science at Columbia University.

Winston Lin is a lecturer and research scholar in the Department of Statistics and Data Science at Yale University.

Claudia Gerber is a projects officer at the International Monetary Fund.