Advance Access publication September 14, 2007

Political Analysis (2008) 16:138–152
doi:10.1093/pan/mpm025

Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches

Donald P. Green
Yale University, Department of Political Science,
77 Prospect Street, New Haven, CT 06520-8209
e-mail: donald.green@yale.edu

Lynn Vavreck
UCLA, Department of Political Science,
4289 Bunche Hall Box 951472, Los Angeles, CA 90095-1472
e-mail: lvavreck@ucla.edu (corresponding author)

Analysts of cluster-randomized field experiments have an array of estimation techniques to choose from. Using Monte Carlo simulation, we evaluate the properties of point estimates and standard errors (SEs) generated by ordinary least squares (OLS) as applied to both individual-level and cluster-level data. We also compare OLS to alternative random effects estimators, such as generalized least squares (GLS). Our simulations assess efficiency across a variety of scenarios involving varying sample sizes and numbers of clusters. Our results confirm that conventional OLS SEs are severely biased downward and that, for all estimators, gains in efficiency come mainly from increasing the number of clusters, not increasing the number of individuals within clusters. We find relatively minor differences across alternative estimation approaches, but GLS seems to enjoy a slight edge in terms of the efficiency of its point estimates and the accuracy of its SEs. We illustrate the application of alternative estimation approaches using a clustered experiment in which *Rock the Vote* TV advertisements were used to encourage young voters in 85 cable TV markets to vote in the 2004 presidential election.

1 Introduction

Recent years have seen rapid growth in the number and sophistication of randomized field experiments in political science. Prior to 2000, only a handful of randomized experiments were conducted outside of laboratory settings (for a review, see Green and Gerber 2002); since 2000, dozens of field experiments have been conducted on topics ranging from the effects of political campaigns on voter turnout (summarized in Green and Gerber 2004) to the effects of international election monitors on election fraud (Hyde 2006). These experimental studies have become increasingly ambitious. Whereas early studies focused on the

Authors' note: We thank *Rock the Vote* for permission to use their public service announcements in this field experiment. The authors are grateful to Alan Gerber for suggestions throughout the design phase of this project. We are also grateful to Dan Kotin and Margaret Coblentz, who worked with cable operators, distributed the advertisements, and assembled the data. We thank Terence Leong for his programming expertise. Replication materials are available on the *Political Analysis* Web site.

effects of a campaign in one area, experiments have become more nuanced, examining the effects of campaign stimuli across a variety of electoral contexts and on different types of voters.

Analysis of field experimental data has at the same time grown more methodologically sophisticated. Early field experiments (e.g., Eldersveld 1956), for example, ignored or mishandled the statistical complications that arise when the treatments received by the experimental groups are not the ones to which they were randomly assigned. Rather than lump untreated members of the assigned treatment group together with members of the control group, recent field experiments have modeled this problem using the logic of instrumental variables regression (Gerber and Green 2000). Contemporary field experiments are also much more attentive to the possibility that treatments administered to one individual might have repercussions for the behavior of other individuals in the same social network (Nickerson 2005).

Finally, political scientists have begun to appreciate the statistical complications that arise when clusters of individuals are assigned to treatment and control groups. As Arceneaux (2005) points out, analyzing such data at the individual level requires the analyst to attend to a variety of issues, most importantly, the fact that the SEs from conventional statistical routines must be corrected to account for the fact that individuals within clusters share unobserved characteristics. Conventional SEs may severely understate the sampling variability associated with estimated treatment effects.

This paper builds on the work of Arceneaux (2005) and parallel work in other disciplines (Arellano 1987; Moulton 1990; Raudenbush 1997; Wooldridge 2003; Varnelli et al. 2004) by addressing a set of practical questions that arise in the analysis of cluster randomized experiments. Decisions about inference methods associated with the clustering of individual observations have been discussed in the literatures of economics, psychology, and medicine for more than a quarter century. As early as 1978, scholars engaged in experimentation noted, "analyses of group randomized trials that ignore clustering are an exercise in self-deception" (Cornfield 1978, 101, as quoted in Arceneaux 2005, 174). Yet, awareness of the statistical complications associated with clustered randomization is far from universal. In his recent review of the literature in economics, Jeffrey Wooldridge writes, "while adjusting for clustering is much more common than it was 10 years ago, inference methods robust to cluster correlation are not used routinely across all relevant settings" (2003, 133). In recent years, some of the most well-known field experiments in political science have misestimated their SEs because clustering was not properly accounted for. For example, Wantchekon's (2003) path-breaking study of village-level assignment of political campaigns in Benin did not account for the fact that his individual survey respondents were embedded in village-level clusters. The experimental study of voter mobilization of Gerber and Green (2000) analyzed data at the individual level but did not properly account for the fact that households were assigned to receive mail, phone calls, or visits and that some of these households consisted of two individuals (see Gerber and Green 2005 for a correction). Our aim here is to increase awareness of clustering and its statistical implications.

This methodological investigation was prompted by a series of practical questions that arose while analyzing an experiment involving the random assignment of *Rock the Vote* (RTV) television advertisements in the days leading up to the 2004 presidential election. As described below, the advertisements, which were designed to encourage young people to vote, were randomly assigned to 85 cable television systems in 12 states. Each cable system comprises several thousand voters, and the entire data set encompasses approximately 850,000 registered voters. Of special interest are the 23,869 voters who are 18 and

19 years of age, for whom this election represents the first federal election in which they are eligible to vote and to whom these ads were specifically addressed. The methodological question is what is the most efficient and reliable way to analyze these data? This question was particularly compelling since our previous mass-media turnout experiments suggested the effects of treatment were likely to be small in magnitude, but not zero (Vavreck and Green 2006).

Focusing on the 18–19 year olds, one could analyze an individual-level data set of 23,869 observations or an aggregated data set in which each cable system is an observation ($n = 85$). One must also choose among estimation techniques. One could estimate the average treatment effect using ordinary least squares (OLS) regression. Alternatively, one could apply a random effects regression model, in which the error term is divided into its individual-level and cluster-level components, as estimated using either generalized least squares (GLS) or maximum likelihood estimation (MLE).

Rather than simply ask whether there were material differences in the *results* produced by these techniques as applied to our data, we conducted a series of Monte Carlo simulations in order to form a clearer sense of how these alternative estimation techniques perform under conditions that are analogous to the kinds of clustered randomized experiments that are currently conducted in political science. We do not claim to be breaking new statistical ground. The large-sample properties of these estimators are the subject of an extensive literature; our aim is to give practitioners a sense of the small sample properties of different estimators so that they can better appreciate what is at stake in the trade-offs among them.

The principal findings from our simulation exercise are as follows.

1. In the presence of cluster-level random effects, conventional regression SEs are biased downward, often producing grossly exaggerated test statistics (Moulton 1990, 334).
2. Consistent with well-known results from other disciplines (Wooldridge 2003), our simulations underscore the fact that in the presence of cluster-level random effects the statistical precision with which cluster-randomized experiments estimate treatment effects is powerfully related to the number of clusters rather than the number of individuals within each cluster.
3. Robust cluster standard errors (RCSEs), which are designed to provide consistent SEs in the face of clustered random assignment, are themselves downwardly biased, particularly when the number of clusters is small.
4. RCSEs are also subject to a fair amount of sampling variability (Angrist and Lavy 2002). Therefore, data analysts should be careful not to report RCSE that, by chance, are smaller than conventional OLS SEs. The larger of the two should be used.
5. In the typical field experimental application, where an outcome measure is regressed on a treatment indicator and a covariate (pretest), individual-level and aggregate-level OLS regression[1] give treatment estimates with very similar statistical properties so long as the number of clusters is greater than 10.
6. The SEs associated with aggregate OLS regression tend to be more reliable than RCSEs from individual-level regression. This result implies that when individual-level and

[1]Our simulations assume that the clusters contain similar numbers of individual level observations. When the number of individuals varies across clusters, analysts must choose whether they want to give each individual equal weight or each cluster equal weight. The corresponding aggregate and individual regressions produce similar treatment estimates so long as the parallel regressions have been weighted to reflect one assumption or the other.

aggregate-level regressions produce different SEs, aggregate-level SEs are probably more trustworthy.

7. Random effects regression, whether GLS or MLE, tend to produce experimental treatment estimates that are slightly more efficient than OLS.
8. The SEs from random effects regression are downwardly biased, particularly when the number of clusters is small. However, GLS SEs appear to be slightly more reliable than RCSEs.

Based on these results, we conclude that in comparison to other estimators, GLS provides the most efficient point estimates while at the same time generating estimated standard errors that are less prone to bias. The advantages of GLS are slight, especially when the number of clusters is greater than 12, but there should nevertheless be a presumption in favor of using GLS to supplement OLS results reported with RCSEs. Applying these estimators to the RTV data, we show how the various approaches play themselves out in our application.

2 Statistical Model

The underlying experimental model consists of an outcome variable (Y), a binary treatment (T), a covariate (X), and a disturbance term (ε). These variables are subscripted $i \in (1, \ldots, N)$ to refer to individuals and $c \in (1, \ldots, G)$ to refer to clusters.

$$Y_{ic} = \alpha + \theta T_{ic} + \beta X_{ic} + \varepsilon_{ic}. \tag{1}$$

The key parameter of interest is θ, the treatment effect. In this context, X might be a pretest and β a measure of the persistence of individual differences. The estimation of β is assumed to be of little substantive interest; we focus therefore on the effect of the experimental intervention and how it is best estimated. One could, of course, omit X entirely from equation (1), as T is assigned at random and therefore is independent of ε. Regardless of whether X is included, one obtains unbiased estimates of θ; the advantage of including X is the potential to reduce disturbance variability, which would result in more precise estimates of θ (Arceneaux 2005).

In this example, the unit of random assignment is assumed to be the cluster. In other words, an entire cluster receives either $T = 0$ or $T = 1$. Thus, it makes sense to consider the possibility that people within the same cluster share similar disturbances. We therefore model the disturbance term as

$$\varepsilon_{ic} = v_c + u_{ic}, \tag{2}$$

where v_c represents the cluster-specific disturbance and u_{ic} the disturbance that is idiosyncratic to a particular individual. Because the treatments are assigned at random, they are by design independent of the random effects (v_c) associated with each cluster.

Random assignment of the treatment ensures that OLS provides unbiased estimates of the treatment effect (θ). However, OLS will generate biased SEs when the variance of $v_c > 0$. The reason is that OLS assumes that $E(\varepsilon_{ic}\varepsilon_{jc}) = 0$ for $i \neq j$, whereas equation (2) implies that subjects within the same cluster will in fact share similar disturbances. Within a given cluster, the u_{ic} will be independent, but individuals will nonetheless share the same v_c, which means that $E(\varepsilon_{ic}\varepsilon_{jc}) > 0$.

This problem may be addressed in three ways. The first approach is to aggregate the data at the cluster level. Each observation represents a cluster-level average. Letting, for

example, $\overline{Y}_c$ denote the average value of Y_{ic} for cluster c, we may write the aggregate-level regression model as

$$\overline{Y}_c = \alpha + \theta T_c + \beta \overline{X}_c + \overline{\varepsilon}_c. \tag{3}$$

This approach resolves the issue of correlated disturbances by eliminating within-cluster variation. Although this approach causes the sample size to diminish from NG to G, the loss of efficiency is to some extent offset by the decline in disturbance variance since the variance of $\overline{\varepsilon}_c$ is generally much lower than the variance of ε_{ic}. The main advantage of this approach is simplicity: working with cluster-level data lends itself to easily interpreted graphical representations of the apparent causal effect of the treatment.

A second approach is to maintain the individual-level structure of the data, using OLS but correcting the resulting SEs. Whereas conventional OLS SEs of the treatment effect take the form $\sqrt{\sigma^2_{u_{ic}}(T'MT)^{-1}}$, where $M = I - X(X'X)^{-1}X'$ and I is an NG $\times$ NG identity matrix, RCSEs take the form

$$\text{RCSE}_\theta = \sqrt{(T'MT)^{-1}T'M\Omega MT'(T'MT)^{-1}}.$$

OLS assumes that $\Omega = \sigma^2_{u_{ic}} I$, which is to say that its off-diagonal elements are all zero. RCSEs, by contrast, impose the following structure on Ω. Its off-diagonal elements are assumed to be zero only for pairs of observations from different clusters. Within clusters, the $\varepsilon_{ic}\varepsilon_{jc}$ element is permitted to be nonzero. Empirically, RCSEs are estimated using the estimated residuals for these within-cluster off-diagonal elements (Williams 2000).[2] Intuitively, the more clusters, the more off-diagonal zero elements, and the more RCSEs resemble conventional OLS SEs. Note that RCSEs, which were inspired by ideas in Huber (1967), are sometimes confused with "robust" Huber-White SEs (White 1980). Both approaches allow for heteroskedasticity, but only RCSEs address the clustered random assignment issue discussed here.

RCSEs are intended to correct for what is usually termed the "design effect" of cluster-randomized experiments (Donner and Klar 2000). Let m denote the average number of observations within each cluster. Let ρ denote the intraclass correlation, which is a nonnegative value defined as

$$\rho \equiv \frac{s^2_{\text{Between}}}{s^2_{\text{Between}} + s^2_{\text{Within}}},$$

where s^2_{Between} refers to the variance among cluster means and s^2_{Within} refers to the average variance within each cluster. The design effect of a cluster-randomized experiment is $1 + (m\text{–}1)\rho$, which is to say that a conventional OLS SE would be biased by a factor of $\sqrt{1 + (m-1)\rho}$. The value of ρ depends on the application. Values of 0.001 seem to be common among medical trials, but higher values are observed in applications where localities or other widely varying groups are the unit of randomization (Donner and Klar 2004). It is difficult to say what values of ρ are common in political science applications as

[2]Unlike conventional SEs, which are the square roots of the diagonals of the $k \times k$ matrix $(\varepsilon'\varepsilon/(N^* - k))(X'X)^{-1}$, RCSEs are the square roots of the diagonal elements of the $k \times k$ matrix $((N^* - 1)/(N^* - k)) \cdot (G/(G-1))(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}$, where k is the number of columns in X, N^* is the total number of observations across all clusters, and G is the number of clusters. In the latter formula, the off-diagonal elements of $\varepsilon\varepsilon'$ are zero for observations from different clusters.

they have not been the objects of systematic study, but even a relatively small value of ρ can generate severe bias when the number of observations in each cluster is large.

For example, suppose an experiment assigned 1001 people in each of 25 clusters to the treatment group and a like number of people in 25 clusters to the control group. For $\rho = 0.003$ and $(m - 1) = 1000$, the design effect would be 4, implying that conventional OLS SEs would be downwardly biased by a factor of 2. In the presence of clustering, tests of statistical significance based on conventional SEs can be grossly misleading (Murray 1998). In this example, the threshold for statistical significance at the 0.05 level using a two-tailed test is ± 3.92 rather than ± 1.96.

Rather than use OLS, which is inefficient in the presence of correlated disturbances, analysts frequently turn to GLS. So-called random effects regression models the cluster-level effects (v_c) as though they were drawn from a normal distribution with finite and constant variance. By partitioning the disturbance variance in this way, random effects regression in effect downplays the influence of outlying clusters. Although the use of random effects regression is sometimes controversial in the context of observational research[3] (Green et al. 2001), this type of model is well suited for use with cluster-randomized experiments because, by design, the random effects associated with each cluster are independent of the assigned treatment. There are different algorithms for generating random effects regression estimates and SEs, GLS, and MLE (Greene 2003, chapter 13). Our simulations address the question of which of the two provides more efficient estimates and reliable SEs, while at the same time gauging the extent to which random effects regression represents an improvement over individual- or aggregate-level OLS.

3 Simulation

The Monte Carlo simulation results were generated using STATA 9/SE and confirmed using GAUSS 6.0. The structure of the simulation reflects an imagined experiment in which a set of pretest scores are observed (X_{ic}), an intervention is randomly administered (T_{ic}), and postintervention outcomes are measured (Y_{ic}). The structural parameters of our simulation have $X_{ic} \sim N(0,4)$, $\theta = 0.5$, and $\beta = 0.85$, but it should be stressed that these values are inconsequential. Because our estimators are each unbiased, the value of θ is immaterial; the focus of the simulation is on the sampling variability of the estimators and the average value of the SEs that each estimator generates. Similarly, changing the dispersion of X_{ic} or the magnitude of β affects the precision with which θ is estimated only if we omit X_{ic} from the model, but all the simulations we report include it. It goes without saying that failure to control for covariates that strongly predict the outcome undercuts the precision with which experimental treatment effects are estimated. The principal reason to include a covariate in our simulations is that doing so provides a comparison between individual-level analysis, which controls for X_{ic}, and aggregate-level analysis, which controls for the mean of X_{ic} within each cluster.

The key parameters in our simulations are the cluster- and individual-level disturbances, which are independently normally distributed with $\text{Var}(v_c) = 0.1$, and $\text{Var}(u_{ic}) = 1.0$.[4]

[3]Random effects regression will generate biased estimates if the cluster-level effects are correlated with the treatment variable. Random assignment procedures allow the experimenter to assume the statistical independence of cluster-level effects and the treatment; nonexperimental researchers are forced to invoke substantive arguments when defending the randomness of the random effects.

[4]Our simulation routines allow for correlation across the disturbances and between the disturbances and the pretest (X). Introducing these correlations does not appear to change the results in any material way. Interestingly, when the individual-level disturbances are correlated with the pretest scores, the individual-level OLS regression becomes less efficient vis-à-vis its aggregate counterpart, although both remain unbiased.

These assumptions imply a value of ρ of 0.1. Making ρ higher accentuates the bias associated with conventional OLS SEs but does not change the relative standing of the estimators we examine.

The variables of most interest for simulation purposes are the number of observations and the number of clusters. We report the results of 10,000 replications of each simulation evaluated at various sample sizes and numbers of clusters. In all our simulations, we place an equal number of individuals into each cluster. (As noted above, researchers facing clusters of unequal size must decide whether they want an average treatment effect at the individual level or at the cluster level and weight the data accordingly.) In each simulation, we "treat" ($T = 1$) exactly half of the clusters.

4 Monte Carlo Results

Table 1 compares the precision with which two OLS regressions, individual-level and aggregate-level, recover the treatment effect. The top panel of Table 1 considers the case in which 480 observations are distributed among clusters, which range in number from 8 to 120. The middle panel repeats this exercise for the case of 1200 observations, and the bottom panel considers the case of 7200 observations. Consistent with the findings of many previous authors (Angrist and Lavy 2002; Donner and Klar 2000; Wooldridge 2003), we find the number of clusters to be much more important than the number of observations. When there are eight clusters, for example, increasing the number of observations from 480 to 7200 scarcely affects the empirical SE of the estimates. At 480 observations, the SE for individual-level analysis is 0.25, as compared to 0.24 for 7200 observations. On the other hand, the number of clusters plays a crucial role in shaping the SE. At 480 observations, for example, increasing the number of clusters from 8 to 80 cuts the SE by more than 50%. The implications for research design are clear: when planning cluster-randomized experiments, strive to maximize the number of clusters, even at the cost of losing some individual observations. Table 1 shows that 480 observations distributed across 120 clusters is far more efficient than 7200 observations packed into 20 or fewer clusters.

Conventional SEs, as noted above, can be severely misleading when applied to cluster-randomized data. Table 1 indicates that this bias is particularly severe when the number of observations per cluster is large. For example, when 7200 observations are packed into 12 clusters, the conventional SE estimate averages 0.025, whereas the true SE is approximately 0.192. In this instance, *t*-values would be off by a factor of seven. Even when 7200 observations are placed in 120 clusters, the average conventional SE is less than half as large as the true SE.

The results presented in Table 1 also illustrate the surprising extent to which individual- and aggregate-level analyses are similar in terms of efficiency. Although intuition might suggest that individual level data—with an individual-level covariate—contain more information than cluster-level aggregates, we find little difference between the two when estimating the treatment effect using OLS. With eight clusters, the SEs differ by less than 10%, and with 40 or more clusters, the two techniques are functionally equivalent. The practical implication of this result is that one gains little by disaggregating experimental units into smaller components in an effort to increase the number of observations.

Finally, Table 1 addresses the question of whether the SEs generated by the robust cluster procedure accurately reflect the true sampling variability associated with the OLS estimates. It is apparent that the RCSEs are biased downward, and this bias is particularly noticeable when the number of clusters is below 20 (Rogers 1993). For example, with 7200 observations in eight clusters, the average RCSE is 0.19, whereas the estimates in fact have

Table 1 Comparison of empirical standard errors and estimated standard errors, for individual and aggregate OLS regressions estimating an experimental treatment effect

Observations	*Clusters*	*Empirical standard error of individual level OLS*[a]	*Average conventional OLS standard error*[b]	*Average estimated robust clustered standard error*[c]	*Empirical standard errors of aggregate OLS*[d]	*Average estimated aggregate standard error*[e]
480	8	0.252	0.101 (0.010)	0.205 (0.069)	0.267	0.253
480	12	0.212	0.099 (0.007)	0.186 (0.047)	0.215	0.209
480	20	0.172	0.097 (0.005)	0.160 (0.029)	0.172	0.171
480	40	0.137	0.096 (0.004)	0.133 (0.016)	0.137	0.137
480	80	0.117	0.096 (0.003)	0.114 (0.009)	0.117	0.116
480	120	0.109	0.096 (0.003)	0.107 (0.007)	0.109	0.108
1200	8	0.243	0.064 (0.006)	0.193 (0.065)	0.264	0.243
1200	12	0.197	0.062 (0.004)	0.173 (0.044)	0.204	0.197
1200	20	0.155	0.062 (0.003)	0.145 (0.027)	0.157	0.155
1200	40	0.118	0.062 (0.002)	0.113 (0.013)	0.118	0.116
1200	80	0.092	0.061 (0.001)	0.090 (0.007)	0.092	0.092
1200	120	0.082	0.061 (0.001)	0.081 (0.005)	0.082	0.082
7200	8	0.239	0.026 (0.002)	0.189 (0.064)	0.250	0.236
7200	12	0.192	0.025 (0.002)	0.167 (0.042)	0.195	0.189
7200	20	0.147	0.025 (0.001)	0.136 (0.024)	0.149	0.146
7200	40	0.104	0.025 (0.000)	0.100 (0.012)	0.104	0.104
7200	80	0.075	0.025 (0.000)	0.074 (0.006)	0.076	0.075
7200	120	0.062	0.025 (0.000)	0.062 (0.004)	0.062	0.063

[a]This is the SD of the estimated treatment effects across 10,000 Monte Carlo replications based on individual-level data.
[b]This is the average OLS SE for the estimated treatment effect across 10,000 Monte Carlo replications based on individual-level data. SDs in parentheses.
[c]This is the average RCSE for the estimated treatment effect across 10,000 Monte Carlo replications based on individual-level data. SDs in parentheses.
[d]This is the SD of the estimated treatment effects across 10,000 Monte Carlo replications based on aggregate-level data.
[e]This is the average OLS SE for the estimated treatment effect across 10,000 Monte Carlo replications based on aggregate-level data.

an empirical SE of 0.24. By contrast, the SEs generated by aggregate OLS regression more closely reflect the true sampling variance of the estimates.

Not only are the RCSEs biased when the number of clusters is small, the RCSEs are themselves subject to considerable sampling error. Our simulations show, for example, that although robust clustered SEs are used to generate an upward correction to the conventional SEs, it is possible for estimated RCSEs in particular samples to be smaller than the naive SE estimates generated by OLS. Those who estimate RCSEs should report conventional SEs instead when the latter prove larger than SEs generated by the robust cluster procedure.

Suppose one decides to use individual-level data to estimate treatment effects. Table 2 addresses the question, to what extent does random effects regression—whether via GLS or MLE—improve the efficiency with which treatment effects are estimated? In theory, random effects regression is more efficient than OLS. In our simulations, however, the efficiency gains turn out to be small, less than 5% and in many cases less than 1%. There is effectively no difference between GLS and MLE. However, both have downwardly biased SEs, with the biases of MLE-generated SEs being more severe.[5] These biased SEs, unfortunately, may give users of MLE random effects regression the misleading impression that this procedure has produced a dramatic gain in efficiency. In actuality, the precision of GLS and MLE is only slightly better than OLS. The main virtue of GLS appears to be that its SEs are less biased than RCSEs.

Taken together, Tables 1 and 2 suggest that GLS provides the most accurate estimates and SEs. The advantage of using GLS is relatively slight. Nevertheless, it is useful to have a victor in the competition among alternative estimation methods. From a broader perspective, the risk of having so many analytic options is that researchers may be tempted to pick the one that seems to produce the most congenial results.

5 Empirical Application: *Rock the Vote* Television Advertisements

Prior to the presidential election in November 2004, we assembled a nationwide list of cable systems that covered only a single zip code. Small cable TV systems are a fertile source of experimental data for social scientists because their small size makes them inexpensive and conducive to large-N randomized studies. In order to test the televised messages in an environment that would not be dominated by other election-related advertisements, we removed all cable systems in 16 states that the *Los Angeles Times* classified as presidential battlegrounds (closely contested states). We then excluded any systems that had no time available in prime time during the week before the election or that cost more than $15 per 30-second advertisement on the USA television network. We excluded all systems in Mississippi because its voter file is very difficult to obtain. This left 85 cable systems for randomization.

Random assignment of the cable systems took place as follows. Each system was matched with one or two other systems in the same state according to its past turnout rate in presidential elections. This procedure resulted in 40 strata containing the 85 cable systems. After sorting the list of 85 cable systems by strata and then by a random number, the first cable system in each stratum was assigned to the treatment condition, the others to

[5]The one virtue of MLE appears to be the smaller sampling variance of its estimated SEs. In terms of mean squared error, MLE's SEs are competitive with other estimators. However, the mean squared error criterion makes most sense in the context of interpreting a single experiment. For purposes of building a cumulative experimental literature, it makes more sense to emphasize unbiasedness, on the grounds that sampling variance will diminish as experimental results accumulate.

Table 2 Comparing the efficiency of random effects regression and OLS, using individual-level data

Observations	*Clusters*	*Empirical standard error of individual level OLS*[a]	*Empirical standard rrror for individual level GLS*[b]	*Empirical standard error for individual level MLE*[c]	*Average estimated robust clustered standard error*[d]	*Average estimated standard error for GLS random effects*[e]	*Average estimated standard error for MLE random effects*[f]
1200	8	0.243	0.237	0.237	0.193 (0.065)	0.223 (0.070)	0.191 (0.058)
1200	12	0.197	0.194	0.194	0.173 (0.044)	0.188 (0.044)	0.170 (0.040)
1200	20	0.155	0.154	0.154	0.145 (0.027)	0.152 (0.026)	0.143 (0.025)
1200	40	0.118	0.117	0.117	0.113 (0.013)	0.116 (0.013)	0.112 (0.013)
1200	80	0.092	0.092	0.092	0.090 (0.007)	0.091 (0.007)	0.090 (0.007)
1200	120	0.082	0.082	0.082	0.081 (0.005)	0.082 (0.005)	0.081 (0.005)
7200	8	0.239	0.225	0.225	0.189 (0.064)	0.215 (0.069)	0.187 (0.055)
7200	12	0.192	0.185	0.185	0.167 (0.042)	0.179 (0.043)	0.163 (0.037)
7200	20	0.147	0.145	0.145	0.136 (0.024)	0.142 (0.024)	0.134 (0.023)
7200	40	0.104	0.103	0.103	0.100 (0.012)	0.102 (0.012)	0.099 (0.012)
7200	80	0.075	0.075	0.075	0.074 (0.006)	0.074 (0.006)	0.073 (0.006)
7200	120	0.062	0.062	0.062	0.062 (0.004)	0.062 (0.004)	0.062 (0.004)

Note. Based on 10,000 repetitions.
[a]This is the SD of the estimated treatment effects across 10,000 Monte Carlo replications, using OLS.
[b]This is the SD of the estimated treatment effects across 10,000 Monte Carlo replications, using GLS random effects regression.
[c]This is the SD of the estimated treatment effects across 10,000 Monte Carlo replications, using MLE random effects regression.
[d]This is the average RCSE for the estimated treatment effect across 10,000 Monte Carlo replications. In parentheses are the SDs of estimated SEs across replications.
[e]This is the average SE for the estimated treatment effect across 10,000 Monte Carlo replications, based on GLS. In parentheses are the SDs of estimated SEs across replications.
[f]This is the average SE for the estimated treatment effect across 10,000 Monte Carlo replications, based on MLE.

control. As expected, an aggregate level regression of treatment on age, gender, a dummy variable for missing gender, and cable system population (including dummy variables for 39 of the 40 matching strata) shows no systematic relationship among these covariates and assignment to treatment condition. An F-test testing the joint significance of these predictors is nonsignificant, $F(4,41) = 0.54$, $p = 0.71$.

People living within the treatment systems saw two different 30-second advertisements produced by *Rock the Vote*. Both advertisements used the same format. The first dealt with the draft and the second, with education. In the draft advertisement, a young couple dancing at a party is talking about the man's new job. He is very excited to be working in promotions and hopes to start his own firm in 6 months. The woman interrupts him and says, "That's if you don't get drafted." The man is puzzled. She clarifies, "Drafted, for the war?" He responds, "Would they do that?" The advertisement closes with everyone at the party looking into the camera and the words, "It's up to you" on the screen. The voiceover says, "The Draft. One of the issues that will be decided this November. Remember to vote on November 2nd." The closing image is of the *Rock the Vote* logo on a black screen.

The second *Rock the Vote* advertisement dealt with education. A young man arrives at work with news that he has been accepted to college. His colleagues congratulate him and one of them asks, "Books, room, board, tuition ... how can you pay for all of that?" The advertisement closes with everyone looking out at the camera and the words, "It's up to you" written on the screen. The voiceover is similar to the one above but with education substituted for draft. We showed both advertisements equally in all cable systems.

Perhaps reflecting its putative effectiveness as a mobilizing tool, "The Draft" advertisement acquired some notoriety during the presidential campaign when the Chairman of the Republican National Committee, Ed Gillespie, wrote a widely publicized letter to *Rock the Vote* President Jehmu Greene demanding that the organization stop talking about the issue of the military draft. Gillespie said in his letter that the draft issue was just an "urban myth" and that the advertisement and its campaign were being conducted with "reckless disregard for the truth." He continued in the letter to take *Rock the Vote* to task for calling itself a nonpartisan 501(C)3 group. Copies of the Chairman's letter were forwarded by the sender to the presidents of major networks, such as NBC, MTV, and HBO, whereupon *Rock the Vote* expanded the "carbon copy" list to include political humorists like Jon Stewart and David Letterman. Although the *Rock the Vote* advertisements became a news story in their own right, they were aired relatively infrequently. *Rock the Vote* does not pay TV stations to air its material; rather, it relies on its corporate partners to place its advertisements for free, as public service announcements. As a result, the advertisements were seldom shown in smaller media markets, where *Rock the Vote's* corporate partners had little control over television programming.

In our experiment, both advertisements were shown as many times as budget constraints would allow. During the last eight days of the election campaign, the advertisements aired four times per night in prime time on each of the following channels: USA, Lifetime, TNT, and TBS.[6] Approximately, two-thirds of American households subscribe to cable TV, and according to the National Cable Commission, these are among the most popular cable channels.

The outcome variable in our study, voter turnout, was obtained from public records of who cast ballots in the 2004 November election. We contracted with Polimetrix, a nonpartisan research firm based in Palo Alto, CA, to assemble validated turnout information for the zip codes in our experiment. Consistent with past voter turnout experiments, the

[6]In systems where time was not available on all these channels, we substituted ESPN for the sold-out network.

Table 3 Estimated treatment effects of *Rock the Vote* advertisements among 18- to 19-year-old voters, using various estimators

	Estimated treatment effect	*Standard error*
OLS: Individual-level data, conventional standard errors	2.4	0.7
OLS: Individual-level data, robust cluster standard errors	2.4	1.4
OLS: Aggregate-level data, conventional standard errors	2.4	1.5
Generalized least squares random effects regression	3.0	1.4
Maximum likelihood estimation: Random effects regression	2.9	1.1

Note. $n = 23{,}869$ people in 85 clusters. Each model includes dummy variables as controls for the 40 strata within which cable systems were randomly assigned. Estimates represent the percentage point increase in turnout associated with the random introduction of *Rock the Vote* advertisements.

turnout rate used here is the proportion of *registered* voters who cast ballots. Turnout among registered 18- and 19-year-olds in our sample was 54%, which is substantially lower than the 64% turnout rate among all voters in our sample.

6 Empirical Results

The model we estimate is below. Recall that randomization of the clusters took place after the clusters were categorized into 40 strata based on past voter turnout. Within each stratum, clusters were assigned to treatment or control. Thus, the model includes 39 dummy variables marking each of the strata (X_k) less one:

$$Y_i = \alpha + \theta T_i + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \varepsilon_i. \quad (4)$$

The key parameter of interest is again θ, the treatment effect. The model does not include individual covariates such as age or gender because these independent variables have almost no predictive power within this narrow band of the voting-age population. Past voting behavior, which is ordinarily a strong predictor in models of this type, is inapplicable to first-time voters.

Table 3 presents the results from a range of different estimation strategies, as applied to 23,869 18–19 year old registered voters distributed across 85 cable system clusters. The coefficients associated with the strata dummy variables have been omitted from the table for ease of presentation. The aggregate-level data were weighted proportional to the number of 18- and 19-year-old registered voters in the cable systems, so that both individual- and aggregate-level estimators have the same estimand, namely the average treatment effect among individuals in clusters assigned to receive the TV advertisements.

The first row of Table 3 shows the results of a naive regression with conventional SEs. These SEs in effect assume that each individual, not cluster, was randomly assigned to treatment or control. If there were no intracluster correlation—as would occur if individuals were randomly assigned to cable systems—these naïve SEs would accurately reflect the expected sampling error associated with the OLS estimate. The second row calls this

assumption into question. The RCSE (1.4) is twice the naive SE (0.7), suggesting that the design effect in this application is approximately 4. The aggregate OLS regression produces a SE that coincides with the RCSE (1.5). All three of these techniques generate the same estimate (2.4), suggesting that turnout was boosted by 2.4 percentage points among those living in cable systems assigned to the treatment group. However, the *t*-ratios vary markedly. The conventional OLS SEs imply a *t*-ratio of 3.4, whereas the RCSEs cause the *t*-ratio to fall to 1.7.

The last two rows of Table 3 present the random effects regression estimates. GLS produces a slightly higher estimate 3.0 with a SE of 1.4. This SE estimate closely coincides with the RCSE, which is to be expected given the large number of clusters. MLE produces an estimate of 2.9 but with a suspiciously low SE of 1.1. Ordinarily, one might be tempted to select the estimate that has the lowest SE, but our simulations suggest that one should be skeptical of the SEs produced by MLE. By the same token, one might be hesitant to select the GLS estimates without a principled basis as they turn out to be the largest in this particular instance. Again, simulations give us some confidence that GLS is the best estimator.

Substantively, the results suggest that age-targeted advertisements can increase turnout in high-salience elections. These effects are impressive in light of the fact that they occur in the context of a presidential election that generated an unusually high rate of voter turnout. They are also impressive when compared to other types of mobilization efforts about which we have experimental data. An average treatment effect of three percentage points among newly registered young people is substantial when viewed against the backdrop of other voter mobilization experiments. These public service announcements, though not personally delivered, were more effective in stimulating turnout than prerecorded phone calls, direct mail, or e-mail (Green and Gerber 2004), three other voter mobilization tactics that are thought to be ineffective because they are impersonal. Thus, the findings presented here suggest that further refinement is needed in the prevailing classification of personal GOTV appeals as effective and impersonal GOTV appeals as ineffective.

7 Conclusion

As political scientists conduct randomized experiments at the level of the household, precinct, or media market, they must attend to the statistical complications that arise when clusters, not individuals, are assigned to experimental groups. The same holds for analysis of observational data; rarely when assessing the causal influence of state- or national-level policy variation do survey analysts take account of the fact that individuals are clustered by state. In large-N studies where the number of clusters is small, design effects are likely to distort conventional OLS SEs.

When analyzing clustered data, both experimental and observational researchers may improve the quality of the inferences they draw from OLS by using RCSEs instead of conventional SEs. We should note that RCSEs should not be confused with "robust" SEs that address heteroskedasticity but do not make use of clustering information. Robust SEs, tellingly, are virtually identical to the conventional SEs reported in Table 3, which is to say, severely biased.

The suggestion that data analysts calculate RCSEs is subject to the caveat that conventional and RCSEs should be compared to ensure that the latter is not smaller than the former. RCSEs are subject to more sampling variability than conventional SEs, and by chance "corrected" SEs may produce misleading *t*-ratios.

The experimental analyst may also go one step further. Random assignment provides a strong justification for the use of GLS because cluster-level effects are statistically independent of the randomly assigned treatment.[7] Our simulations indicate that random effects regression provides relatively modest gains in terms of efficiency, but gains nevertheless. Moreover, the SEs produced by GLS appear to be slightly more reliable than their RCSE counterparts. The finding that GLS outperforms other estimators enables experimental researchers to commit to using GLS prior to the implementation of an experimental design. Commitments of this type help ensure that the selection of estimators is not endogenous to the estimation process.

Funding

Carnegie Corporation (D05015); Yale Institution for Social and Policy Studies.

References

Angrist, Joshua D., and Victor Lavy. 2002. The effect of high school matriculation wards: Evidence from randomized trials. *National Bureau of Economic Research* Working Paper 9389.

Arceneaux, Kevin. 2005. Using cluster randomized field experiments to study voting behavior. *The Annals of The American Academy of Political and Social Science* 601:169–79.

Arellano, Manuel. 1987. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* 49:431.

Cornfield, J. 1978. Randomization by group: a formal analysis. *American Journal of Epidemiology* 108:100–2.

Donner, Allan, and Neil Klar. 2000. *Design and analysis of cluster randomization trials in health research.* London: Arnold.

———. 2004. Pitfalls of and controversies in cluster randomized trials. *American Journal of Public Health* 94:416–22.

Eldersveld, Samuel J. 1956. Experimental propaganda techniques and voting behavior. *American Political Science Review* 50:154–65.

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* 94:653–63.

———. 2005. Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005). *American Political Science Review* 99:301–13.

Green, Donald P., and Alan S. Gerber. 2002. Reclaiming the experimental tradition in political science. In *Political science: The state of the discipline*, ed. Helen V. Milner and Ira Katznelson, 805–32. 3rd ed. New York: W.W. Norton & Co.

———. 2004. *Get out the vote!: How to increase voter turnout.* Washington, DC: Brookings Institution Press.

Green, Donald P., Soo Yeon Kim, and David Yoon. 2001. Dirty pool. *International Organization* 55:441–68.

Greene, William H. 2003. *Econometric analysis.* 5th ed. Upper Saddle River, NJ: Prentice-Hall.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1:223–33.

Hyde, Susan D. 2006. Foreign democracy promotion, norm development and democratization: Explaining the causes and consequences of internationally monitored elections. Unpublished doctoral thesis, Department of Political Science, University of California, San Diego.

Moulton, Brent. 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* 72:334.

Murray, D. M. 1998. *Design and analysis of group-randomized trials.* New York: Oxford University Press.

Nickerson, David Warwick. 2005. Measuring interpersonal influence. Unpublished doctoral thesis, Department of Political Science, Yale University.

Raudenbush, Stephen W. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 2:173–85.

[7]Although there is no guarantee that cluster-level effects are drawn from a normal distribution, normality is something that can be assessed empirically by plotting the distribution of cluster-level residuals.

Rogers, W. H. 1993. Regression standard errors in clustered samples. *Stata Technical Bulletin* 13:19–23.
Varnelli, Sherri P., David M. Murray, Jessica B. Jenega, and Jonathan L. Blitstein. 2004. Design and analysis of group-randomized trials: A review of recent practices. *Evaluation Methods and Practice* 94:393–9.
Vavreck, Lynn, and Donald P. Green. 2006. Mobilizing voters through TV public service announcements: A field experiment. Unpublished report prepared for the Carnegie Corporation and presented at the MWPSA meeting, Chicago, IL.
Wantchekon, Leonard. 2003. Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics* 55:399–422.
White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48:817–30.
Williams, Rick L. 2000. A note on robust variance estimation for cluster-correlated data. *Biometrics* 56:645–6.
Wooldridge, Jeffrey M. 2003. Cluster-sample methods in applied econometrics. *American Economic Review* 93:133–8.