The Analysis of Experimental Data: Comparing Techniques

Thad Dunning
Department of Political Science
Yale University
P.O. Box 208301
New Haven, CT 06520-8301
thad.dunning@yale.edu

Susan Hyde
Department of Political Science
Yale University
P.O. Box 208301
New Haven, CT 06520-8301
susan.hyde@yale.edu

We are grateful to David Freedman and Don Green for very helpful suggestions. A previous version of this paper was presented at the annual meeting of the American Political Science Association, Boston, Massachusetts, August 28-31, 2008.

Abstract

Experimental data can be analyzed using the intention-to-treat principle, in which units randomly assigned to receive treatment are compared to those assigned to control. In the presence of crossover – for example, when units assigned to the treatment group are instead subjected to the control regime – intention-to-treat analysis usually gives a conservative estimate for the effect of treatment. Analysts may therefore wish to estimate the effect of treatment on the treated, that is, the differential effect of treatment for units who would receive the treatment if assigned to treatment and the control if assigned to control. The point we make in this article is that such estimates are model-dependent, and different models can give very different answers. Issues of experimental design can also substantially complicate estimation of the effect of treatment on the treated. Intention-to-treat analysis, on the other hand, is the most robust way to analyze experiments; in many contexts, the intention-to-treat parameter may also have the most policy as well as social-scientific relevance. We illustrate these points using data from a field experiment in which election monitors were randomly assigned to villages in Indonesia.

1 Introduction

Randomized controlled experiments are increasingly used in political science and related fields.[1] The main attraction of experiments is that they solve pervasive problems of confounding and selection bias. In non-experimental studies, groups of units exposed to an intervention or treatment may be compared to a group of unexposed or control units. However, the groups may be unlike in ways besides exposure to treatment, and these pre-exposure differences may be responsible for post-intervention differences across groups as well.

With experiments, by contrast, random assignment ensures that treated and untreated groups are equivalent prior to the intervention, up to random error. With a large enough number of units, random error will play only a small role. Post-intervention differences across the treatment and control groups can then be reliably attributed to the effect of treatment.

A second attraction of experiments, relative to observational studies, is simplicity and transparency. Experiments can always be analyzed according to the intention-to-treat principle, which measures the causal effect of assignment to treatment or control. With intention-to-treat analysis, few complicated adjustments to the data are typically necessary, and few assumptions need to be invoked to draw causal inferences.

Nonetheless, in some settings analysts may believe that the causal effect of assignment to treatment is not the most relevant parameter. One common source of difficulty in experiments is crossover: experimental subjects randomly assigned to treatment may be subjected to the control regime, while those assigned to control may instead receive the treatment. With crossover, intention-to-treat analysis usually gives a conservative estimate of the effect of treatment.

Analysts may therefore wish to estimate the effect of treatment on the treated, that is, the differential effect of treatment for units who would receive the treatment if assigned to treatment and the control if assigned to control. (As discussed below, this estimand is sometimes called the

[1] See Gerber and Green (2006) for discussions of the use of experimental methods in political science.

effect of treatment on compliers). However, crossover raises issues akin to those raised by observational studies, since confounding variables may be related to receipt of treatment, as opposed to treatment assignment. Correcting for crossover may involve complicated adjustments to the data, and choices are going to be involved in the formulation of the correction.

The point we make in this article is that corrections are model-dependent, and different models can give very different answers. Common techniques for analyzing experiments with block randomization, such as instrumental-variables regression with block "fixed effects," can give estimates that are misleading. In addition, issues of experimental design can substantially complicate estimation of the effect of treatment on the treated. Where feasible, it is preferable to seek to boost compliance through appropriate experimental design modifications, rather than through ex-post adjustment to the data. Intention-to-treat analysis is the most robust way to analyze experiments, because it involves the purest experimental comparison: that is, between groups randomly assigned to treatment and groups randomly assigned to control. In many contexts, the intention-to-treat parameter may have the most policy as well as social-scientific relevance.

We illustrate these points using data from a field experiment in Indonesia, in which international election monitors were randomly assigned to observe local polling places during the 2004 presidential election (Hyde 2008). In this experiment, there was substantial crossover from treatment to control – many locales assigned to election monitoring were not, in fact, visited by monitors – while there was also some crossover from control to treatment – some villages assigned to control were mistakenly visited. There were other technical complications. Random assignment of observers took place within blocks, and the blocks were of different sizes. Treatment effects were heterogenous across blocks as well as related to block size, and the degree of crossover varied from block to block. These features of the Indonesia field experiment motivate several adjustments to the data. As we show in this article, results can be quite sensitive to the choice of technique.

2 A field experiment on election monitoring

The 2004 election was the first direct presidential election in Indonesian history. During the second round runoff, when the incumbent candidate Megawati Sukarnoputri faced her leading challenger Susilo Bambang Yudhoyono, international monitors associated with the Carter Center were randomly assigned to observe local units (villages or neighborhoods) in which polling places were located. The random assignment of election monitors permits the study of several interesting questions (Hyde 2008). For instance, did election monitoring increase or decrease the incumbent's vote share?

With approximately 155 million eligible voters, 17,508 islands, and nearly 580,000 polling stations, randomizing election observers in Indonesia was no small affair. The complexity of the election monitoring task influenced the design of the experiment in several ways.

First, many areas of the country were inaccessible to international observers on election day. The universe of the study is not all villages or polling places in Indonesia but rather a set of villages located within selected districts. In Indonesia, districts are administrative units smaller than provinces but larger than villages.[2]

Second, observer teams were assigned by the Carter Center to particular districts, and random assignment of villages to election monitoring took place within these districts. Randomization took place at the village level, because no complete list of polling stations was available from the central government prior to the election. In all, 20 districts comprising 2165 villages were selected for inclusion in the experiment, with an average of 109 villages per district. Below, we often refer to districts as "blocks."

Third, within villages assigned to monitoring, observers were to select polling stations at random. While the average village had about 18 polling places, on average observers visited

[2]There are five administrative levels relevant to elections in Indonesia: provinces, districts, sub-districts, villages or neighborhoods, and polling stations.

about 1.4 polling places per monitored village.[3] Thus, in this experiment villages were assigned to treatment or control, where treatment means that some (though not all) polling places in the villages were to be monitored.

Finally, the number of villages in the treatment group was substantially larger than the number of villages that observers expected to be able to visit.[4] In consequence, many villages assigned to monitoring were not, in fact, monitored. Across all 20 blocks (districts), just 19.7 percent of the 482 assigned-to-monitoring villages were actually monitored. This rate varied substantially across blocks, ranging from a low of 1.5% in West Java to a high of 73.3% in East Java. In addition, around 1.1% of villages assigned to the control were mistakenly monitored.[5] As we discuss below, these features of the experimental design were far from optimal, and they resulted in substantial crossover.

The features of this experiment present the researcher with several analytic options. First, the outcome variable must be identified: are we interested in the causal effect of monitoring on the number of votes for Megawati in the study population? Or instead a measure of vote share, such as the number of votes for Megawati divided by the number of registered voters?[6]

Second, the analyst must reach a decision on weighting. In this experiment, villages were randomized to treatment and control within blocks (that is, districts). The blocks are of different sizes, and there is reason to suspect that the effect of monitoring is related to block size. Thus, estimated treatment effects for each block must be weighted across blocks to arrive at a global estimator of the effect of monitoring (or assignment to monitoring, as discussed below). One natural approach is to estimate the effect of monitoring on votes for Megawati, by village, within

[3]The average number of registered voters in each village was 5,382. Indonesian law limits polling places to 300 registered voters, implying that on average each village had about 18 polling places.

[4]The Carter Center had 24 observer teams, each comprised of approximately two observers, and polling places were only open from 7 AM to 1 PM in each village.

[5]Some observer teams mistakenly monitored polling stations in villages or neighborhoods assigned to the control regime, that were located near the border between urban neighborhoods; other teams encountered logistical difficulties that caused them to choose to visit villages outside of their assigned list.

[6]We might instead use a measure of vote share such as the number of votes divided by the number of ballots cast (see Hyde 2008).

each block. The number of villages differs across blocks, however, so we must weight each block-level effect by the number of villages in the block. The weighted sum of the block-level effects gives the global effect of monitoring on votes for Megawati in the study population; dividing this sum by the total number of registered voters gives the effect of monitoring on Megawati's vote share.[7]

Finally, as mentioned above, an important feature of this experiment is that there was substantial crossover: in many districts, only a small fraction of the assigned-to-monitoring villages were actually monitored, while some villages assigned to the control regime were also mistakenly monitored. While the experiment can be analyzed according to the intention-to-treat principle, this is likely to give a conservative estimate of the effect of treatment. We may therefore want to estimate the effect of treatment on the treated, that is, on compliers: villages that were actually monitored if assigned to treatment but otherwise were subjected to the control regime. This leads to a choice of models for adjustment.

In the rest of this article, we discuss intention-to-treat analysis and estimation of the effect of treatment on the treated. For both estimators, we use the number of villages per block as our measure of block size, for purposes of weighting treatment effect estimates across blocks. We then compare these estimators to a common but misleading strategy for analyzing such an experiment with block randomization, namely, instrumental variables least squares (IVLS) regression with dummy variables or "fixed effects" for each block. We also discuss alternative, lower-variance estimators where we weight within and across blocks by the number of registered voters, rather than the number of villages.

We then turn to some more foundational issues. First, we point out that different experimental design choices could have resolved some of the analytic issues we discuss in this paper, while at the same time possibly raising others. Then we also raise a philosophy-of-inference ques-

[7] As we describe below, estimators based on this weighting strategy may be unbiased but inefficient. We therefore discuss alternative, lower-variance estimators below.

tion, asking what is the most relevant parameter to estimate: intention-to-treat or the effect of treatment on the treated? In many contexts, we suggest, the intention-to-treat parameter may be most relevant.

3 Intention-to-Treat Analysis

Intention-to-treat analysis relies on a simple principle: in a randomized controlled experiment, units assigned to the treatment group are on average like those assigned to the control group, prior to the intervention. Three parameters are typically of interest in experiments: (a) the average response, if all units were assigned to treatment; (b) the average response, if all units were assigned to control; and (c), the difference between (a) and (b). Parameter (c) is the intention-to-treat parameter (which is sometimes called the average causal effect or average treatment effect).

With intention-to-treat analysis, researchers disregard who actually receives treatment. It can be misleading to compare units that actually receive treatment to the rest of the experimental population.[8] However, randomization ensures that the average response of units assigned to treatment is an unbiased estimate for (a), and the average response of units assigned to control is an unbiased estimate of (b). Thus, the difference between these two averages is an unbiased estimate for (c). For further discussion of the intention-to-treat principle, see Freedman (2006).

Here, the intention-to-treat estimator for a given block is votes for Megawati in villages assigned to monitoring, divided by the number of assigned-to-monitoring villages, minus votes for Megawati in villages assigned to the control, divided by the number of assigned-to-control villages. Within a block, in other words, the intention-to-treat estimator is the simple average of votes for Megawati by village, in the assigned-to-monitoring group, minus the simple average of

[8]That is, it may be misleading to compare those who receive treatment to the rest of the experimental population, including both units originally assigned to control and units assigned to treatment who got the control. In mammography trials, for example, there is often substantial noncompliance with the experimental protocol; women who opt to get screened for breast cancer may be unlike those who do not, in ways that matter for health outcomes (see Freedman, Petitti, and Robins 2004).

Megawati's votes by village, in the assigned-to-control group.

We must then weight the block-by-block intention-to-treat estimators to arrive at a global intention-to-treat estimator for the study population. Here, because blocks have different numbers of villages, the estimated number of votes gained or lost for Megawati per village should be weighted by the number of villages in each block. For instance, the weighted sum of the block-level effects estimates the total number of votes gained or lost due to treatment assignment.

In symbols, for each block j, let M_{ij}^{T} be votes for Megawati in assigned-to-monitoring village $i = 1, ..., T_j$ and M_{ij}^{C} be votes for Megawati in assigned-to-control village $i = 1, ..., C_j$. Here, T_j is the number of villages assigned to treatment in block j, and C_j is the number of villages assigned to control.

The intention-to-treat estimator in block j is then

$$\text{ITT}_j = \frac{\sum_{i=1}^{T_j} M_{ij}^{T}}{T_j} - \frac{\sum_{i=1}^{C_j} M_{ij}^{C}}{C_j}. \tag{1}$$

Equation (1) estimates the difference between the average number of votes for Megawati by village, if all villages in block j were assigned to treatment, and the average number of votes for Megawati by village, if all villages in the block were assigned to control. Notice that T_j and C_j are fixed, not random, numbers for each block.[9]

Because blocks have different numbers of villages, we need a weighted sum of the block-level effects on votes for Megawati per village, where the weights are the numbers of villages. The block-level effect for block j is given by equation (1). Thus, the total effect of assignment to monitoring on votes for the incumbent is

$$\text{ITT}_{\text{total votes}} = \sum_{j=1}^{J} (T_j + C_j)(\text{ITT}_j), \tag{2}$$

[9]The importance of this observation is that, unlike other estimators discussed below, equation (1) is not subject to ratio-estimator bias.

where J is the number of blocks. We can also estimate the global effect of assignment to monitoring on votes for Megawati by village as

$$\text{ITT}_{\text{votes per village}} = \frac{\text{ITT}_{\text{total votes}}}{\sum_{j=1}^{J}(T_j + C_j)}. \tag{3}$$

Finally, the global effect of assignment to monitoring on Megawati's vote share in the study population is

$$\text{ITT}_{\text{vote share}} = \frac{\text{ITT}_{\text{total votes}}}{\sum_{j=1}^{J} reg_j}, \tag{4}$$

where reg_j is the number of registered voters in block j.

Estimation of the variance of these estimators is straightforward. For instance, using equation (2), the estimated variance of the intention-to-treat estimator for each block j is

$$\widehat{\text{var}}(\text{ITT}_j) = \frac{1}{T_j}\frac{1}{T_j - 1}\sum_{i=1}^{T_j}\left(M_{ij}^T - \overline{M_j^T}\right)^2 + \frac{1}{C_j}\frac{1}{C_j - 1}\sum_{i=1}^{C_j}\left(M_{ij}^C - \overline{M_j^C}\right)^2. \tag{5}$$

Here, $\overline{M_j^T}$ is the mean vote for Megawati in the assigned-to-monitoring villages in block j, and $\overline{M_j^C}$ is the mean vote for Megawati in the assigned-to-control villages in the block.[10] The estimated variance of the global ITT estimator for votes gained or lost by Megawati across all blocks is then

$$\begin{aligned}\widehat{\text{var}}(\text{ITT}_{\text{total votes}}) &= \widehat{\text{var}}\Big(\sum_{i=1}^{J}(T_j + C_j)\text{ITT}_j\Big) \\ &= \sum_{i=1}^{J}(T_j + C_j)^2\widehat{\text{var}}(\text{ITT}_j)\end{aligned} \tag{6}$$

The standard error is the square root of (6).[11] Variances for equations (3) and (4) are calculated

[10] With random assignment to treatment and control groups, it is valid to calculate the variance of the difference of means, (ITT_j), as if we had two independent samples, as in equation (5); see, e.g., Freedman, Pisani, and Purves (1998: 510-13).

[11] Equation (6) is valid because the ITT effects are independent across blocks (randomization took place within blocks). Thus, the variance of the sum is the sum of the variances.

analogously.[12]

Table 1 reports ITT_j, the intention-to-treat estimator for each block (that is, each district). The final rows of the table then report the three global estimators in equations (2), (3), and (4). According to the point estimates obtained through this weighting scheme (see section 6 for an alternative), assignment to monitoring garnered Megawati 223,650 extra votes, or about 103 votes per village, and raised her vote share by 0.019. With estimated standard errors of about 146, 653, 68, and 0.013, respectively, however, the intention-to-treat estimates are not significantly different from zero.

4 Effect of Treatment on the Treated

As mentioned above, there was substantial crossover in this experiment. On average, only 19.7 percent of the villages assigned to monitoring were in fact monitored. In addition, about 1.1 percent of the assigned-to-control villages were mistakenly monitored by election observers. With a high degree of crossover from the treatment to the control arm, intention-to-treat analysis is likely to give a conservative estimate of the effect of treatment on the treated. After all, only a relatively small fraction of the villages assigned to the treatment group were actually exposed to the treatment.

In this section, we discuss estimation of the effect of treatment on the treated (ETT), that is, the differential effect of treatment on villages that are monitored if assigned to monitoring but

[12]For equation (3), the formula is

$$\widehat{\text{var}}(\text{ITT}_{\text{votespervillage}}) = \frac{\text{var}(\widehat{\text{ITT}}_{\text{total votes}})}{\left(\sum_{j=1}^{J} T_j + C_j\right)^2}.$$

Similarly, for equation (4), the formula is

$$\text{var}(\widehat{\text{ITT}}_{\text{voteshare}}) = \frac{\text{var}(\widehat{\text{ITT}}_{\text{total votes}})}{\left(\sum_{j=1}^{J} \text{reg}_\text{j}\right)^2},$$

where reg_j is the number of registered voters in block j.

Table 1: Votes for Megawati: Estimated Intention-to-Treat (ITT) Effects, by Block

Block (district)	Number of villages (T_j, C_j)	ITT (votes gained or lost per village)
1. Kota Banda Aceh	90 (19, 71)	-0.9
2. Kota Surabaya	163 (34, 129)	899.8
3. Kota Mataram	23 (11, 12)	206.0
4. Sampang	186 (41, 145)	166.8
5. Tabanan	117 (27, 90)	186.0
6. Situbondo	136 (32, 104)	573.2
7. Kota Yogyakarta	45 (20, 25)	96.8
8. Kota Kediri	46 (15, 31)	682.6
9. Kota Medan	156 (30, 126)	-29.3
10. Kampar and Kota Pekan Baru	243 (53, 190)	30.2
11. Kota Samarinda	42 (8, 34)	-132.4
12. Cianjur	343 (68, 275)	-44.1
13. Kota Palangka Raya	31 (6, 35)	-203.6
14. Kota Pontianak	24 (9, 15)	-1,162.4
15. Kota Padang	103 (20, 83)	95.6
16. Palembang	103 (20, 83)	-187.1
17. Kota Bitung	60 (12, 48)	-120.3
18. Kota Ternate	52 (11, 41)	32.2
19. Kota Ambon	56 (18, 38)	-636.2
20. Kota Makassar	146 (28, 118)	58.4
$ITT_{\text{total votes}}$ (s.e.)	2165 (482, 1683)	223,649.6 (146,652.5)
$ITT_{\text{votes per village}}$ (s.e.)	–	103.3 (67.7)
$ITT_{\text{vote share}}$ (s.e.)	–	0.019 (0.013)

that are not monitored if assigned to control.[13] In medical trials, such experimental units are called "compliers;" thus, what we term the effect of treatment on the treated is sometimes called the effect of treatment on compliers.[14]

[13]Sometimes, the effect of treatment on the treated denotes the effect of treatment on both compliers and "always treats" (Freedman 2006), that is, units that receive treatment whether assigned to treatment or control.

[14]When there is one treatment condition and one control condition, the effect of treatment on compliers corresponds to what Imbens and Angrist (1994) call the Local Average Treatment Effect (LATE) (see also Angrist, Imbens and

The estimation of the ETT in this experiment raises several analytic issues. First, randomization occurred within blocks, so we must calculate the ETT separately for each block, just as we did for the ITT above. The global estimator is then a weighted average of the block-by-block estimators, as above. In this section, we use the number of villages as weights, just as we did for the ITT; below, we discuss an alternate weighting strategy.

Second, while there was high crossover on average, there was also substantial variance across blocks; the percent of assigned-to-monitoring villages that were actually monitored runs from 1.5% to 73.3% across blocks. The low rate of compliance in some blocks implies that the ITT and the ETT diverge sharply in these blocks. Finally, both treatment effects and the rate of compliance appear related to block size. As we show below, these features of the experiment lead to marked contrasts between the global ITT and ETT estimators, suggesting that substantive results may be strongly influenced by the choice of analytic technique.

4.1 The ETT estimator

Within a given block, the ETT estimator is the intention-to-treat estimator, divided by the fraction of assigned-to-treatment villages that were actually monitored minus the fraction of assigned-to-control villages that were monitored. In symbols, the effect of treatment on the treated in block j is

$$\mathrm{ETT}_j = \frac{\mathrm{ITT}_j}{\alpha_j - \beta_j} \tag{7}$$

Here, α_j is the fraction of assigned-to-monitoring villages that are actually monitored in block j, while β_j is the fraction of assigned-to-control villages that are mistakenly monitored in that block. The intention-to-treat estimator ITT_j in the numerator of equation (7) is given by equation (1). Equation (7) estimates the effect of monitoring on votes for Megawati per village in block j, for villages that are monitored if assigned to monitoring and not monitored if assigned to control. For

Rubin 1996; Angrist and Krueger 2001). In other contexts, the ETT and LATE may diverge.

further discussion of this estimator, see Freedman (2006).

Now, we must weight the ETT by the number of villages in the block to arrive at a global ETT estimator. For instance, the effect on votes for Megawati in the study population is

$$\text{ETT}_{\text{total votes}} = \sum_{j=1}^{J}(T_j + C_j)(\text{ETT}_j), \tag{8}$$

where J is the number of blocks. We can also estimate the global effect of monitoring on votes for Megawati by village, for compliers, as

$$\text{ETT}_{\text{votes per village}} = \frac{\text{ETT}_{\text{total votes}}}{\sum_{j=1}^{J}(T_j + C_j)}. \tag{9}$$

Finally, the global effect of assignment to monitoring on Megawati's vote share in the study population is

$$\text{ETT}_{\text{vote share}} = \frac{\text{ETT}_{\text{total votes}}}{\sum_{j=1}^{J} reg_j}, \tag{10}$$

where reg_j is the number of registered voters in block j.

Table 2 reports the block-by-block estimator ETT_j for each of the 20 blocks, as well as the global estimators given by equations (8), (9), and (10). The global ETT estimators sharply diverge from the global ITT estimators reported in Table 1. For instance, while the total effect of assignment to monitoring on votes for the incumbent is estimated at 223,649.6 votes, the ETT for votes for Megawati in the study population is -2,717,560. In this experiment, in other words, the point estimate of the global ITT is very large and positive, while the global ETT is very large in absolute value and negative.

Why do the global ETT estimators in equations (8), (9), and (10) differ in sign from the global ITT estimators in equations (2), (3), and (4)? Within a block, the ITT and the ETT estimators must have the same sign: see, for instance, equation (7).[15] In this experiment, however, treatment

[15]This assumes that the fraction of villages monitored in the assigned-to-monitoring group is larger than the fraction

Table 2: Votes for Megawati: Estimated Effects of Treatment on the Treated (ETT), by Block

Block (district)	α_j, β_j	ETT (votes gained or lost per village)
1. Kota Banda Aceh	$\frac{3}{19}, \frac{1}{71}$	-6.0
2. Kota Surabaya	$\frac{3}{34}, \frac{0}{129}$	10,198.2
3. Kota Mataram	$\frac{4}{11}, \frac{0}{12}$	566.6
4. Sampang	$\frac{5}{41}, \frac{0}{145}$	1,367.9
5. Tabanan	$\frac{6}{27}, \frac{0}{90}$	837.1
6. Situbondo	$\frac{10}{32}, \frac{0}{104}$	1834.4
7. Kota Yogyakarta	$\frac{4}{20}, \frac{0}{25}$	484.1
8. Kota Kediri	$\frac{11}{15}, \frac{0}{31}$	930.8
9. Kota Medan	$\frac{5}{30}, \frac{2}{126}$	-194.1
10. Kampar and Kota Pekan Baru	$\frac{5}{53}, \frac{0}{190}$	320.1
11. Kota Samarinda	$\frac{4}{8}, \frac{2}{34}$	-300.1
12. Cianjur	$\frac{1}{68}, \frac{3}{275}$	-11,617.0
13. Kota Palangka Raya	$\frac{1}{6}, \frac{4}{35}$	-30,541.0
14. Kota Pontianak	$\frac{4}{9}, \frac{3}{15}$	-4,755.5
15. Kota Padang	$\frac{4}{20}, \frac{1}{83}$	508.6
16. Palembang	$\frac{11}{20}, \frac{1}{83}$	-347.8
17. Kota Bitung	$\frac{3}{12}, \frac{4}{48}$	-721.8
18. Kota Ternate	$\frac{5}{11}, \frac{1}{41}$	74.8
19. Kota Ambon	$\frac{4}{18}, \frac{1}{38}$	-3,247.6
20. Kota Makassar	$\frac{2}{28}, \frac{2}{118}$	1,072.0
$\text{ETT}_{\text{total votes}}$ (s.e.)	$\frac{95}{492}, \frac{25}{1683}$	-2,717,560 (5,040,584)
$\text{ETT}_{\text{votes per village}}$ (s.e.)	–	-1,255.2 (2,328.2)
$\text{ETT}_{\text{vote share}}$ (s.e.)	–	-0.233 (0.433)

effects are heterogenous across blocks: in some blocks, the ITT and ETT estimators are positive, while in others they are negative (see Tables 3 and 4.1). Moreover, some large blocks have negative but relatively moderate ITT estimates, while they have very negative ETT estimates; this can occur

of villages monitored in the assigned-to-control group, as it is in this experiment: that is, $\alpha_j > \beta_j \; \forall j$.

when the fraction of villages that are monitored in the assigned-to-monitoring group is very low.[16] Not only are the ETT estimates very negative for such blocks, but such blocks receive a large weight in the global estimators, due to their size. Thus, while the weighted averages of the ITT estimators are positive, the weighted averages of the ETT estimators are negative.

Note that the ETT estimator in equation (7) is a ratio estimator and is subject to ratio-estimator bias, because the numerator and denominator of the estimator are both random: for instance, the number of villages actually monitored in the assigned-to-monitoring villages, and the number monitored in the assigned-to-control villages, are random variables. See Freedman (2006: 709) for discussion. Here, the degree of bias is likely to be small, and more appreciable for the assigned-to-treatment villages than the assigned-to-control villages.

Aside from bias, ratio estimators raise issues for the estimation of variance. With a large number of units, the delta method can be tried; however, here the number of villages in each block is often quite small.[17] We obtained jackknifed estimates for the variances, but the estimates are unreliable, due to the small sample size and the low contact rate.[18] The standard errors for the global ETT estimators reported in Table 4.1 are obtained by instrumental-variables regression within blocks; nominal variances for the coefficient estimates are then pooled across blocks using formulas akin to equation (6).[19] The asymptotics for the nominal variances do not necessarily hold,

[16]For instance, Table 1 reports an ITT estimate of -44.1 votes for block 12 (Cianjur), while Table 2 reports an ETT estimate of -11, 617 for the same block. This in turn occurs because the fraction of assigned-to-monitoring villages that are actually monitored is very low: $\alpha = \frac{1}{68}$ or about 0.0147, while $\beta = \frac{3}{275}$ or about 0.0109. The denominator of equation (7) is thus 0.0038 for this block.

[17]The delta method relies on a first-order Taylor series approximation; see Freedman (2008b).

[18]For the jackknife, within each block, we dropped the ith village and calculated equation (7) to get the ith pseudoreplicate; we did this for $i = 1, ..., n$. The "jackknife variance" is the sum of the squared deviations from the mean of the n pseudoreplicates, multiplied by $\frac{n}{n-1}$ and by a finite-sample correction factor $(1 - f)$, where f is the sampling fraction (the fraction of villages assigned to monitoring). Most of the variance in the ETT estimator comes from the treatment group, which typically has far fewer villages than the control group; for purposes of jackknifing the variance, votes for Megawati and the fraction of monitored villages in the control group were held constant at their sample values. Within-block variances were then pooled across blocks, using a formula similar to equation (6). For discussion of the jackknife, see Freedman (2000).

[19]Within blocks, we regress votes for Megawati on an intercept and a dummy variable that equals one if the village was monitored; we instrument by assignment-to-monitoring. Within blocks, the coefficient of this instrumental-variables least squares regression is the within-block ETT estimator; see section 5.

however, and other issues arise (see Freedman 2008a).

4.2 Pooling into "superblocks"

The global ETT estimates reported in Table 2 are highly sensitive to outliers. For instance, if we drop only block 2 from the data set, the global ETT estimates remain highly negative; if we drop only block 12, the global ETT estimates are highly positive, while if we drop both blocks 2 and 12, the global ETT estimates are again large and negative. In addition, the variance of the within-block estimators, and thus of the global estimators, is very large.

This sensitivity to outliers and large variance of estimators is due to the experimental design. In this experiment, there were several very large blocks in which very few assigned-to-monitoring villages were actually monitored; for such blocks, for instance blocks 2, 12, and 13, the denominator of equation (7) is very small, relative to the numerator (that is, relative to the ITT for that block), and thus the absolute value of the ETT is large. These blocks are influential not only because the ETT is large in absolute value, but also because the blocks are large: thus, they receive large weights in the global ETT estimators given by equations (8), (9), and (10). The fact that the number of villages assigned to monitoring and especially the fraction of such villages that were actually monitored is very small, in some blocks, boosts the variance of the within-block ETT estimators.

These features of the experimental design were far from optimal from the point of view of causal inference. Force was dispersed all over the map: each team of electoral observers could only visit several villages on election day, and generally one team was assigned to each block. Among other effects, this greatly increased the variance of ETT estimates within blocks. Moreover, some teams of observers visited numerous control villages, due to logistical and transportation issues, or because they were resistant to the idea of only visiting villages in the assigned-to-monitoring group. We further discuss features of the experimental design below.

For illustrative purposes, we create a new data set by pooling blocks into larger "superblocks," which will limit the sensitivity of our analyses to outliers; we will use this new data set for purposes of comparing the ETT with a different estimator below. We selected blocks to pool by focusing on blocks with low α_j and then merged blocks with similar average values of votes for Megawati by village. This pooling process created ten "superblocks," each with substantially higher values of α_j (and lower values of β_j) than in the previous analysis. Using this new pooled data set, we then recalculated the various ETT estimators in equations (7), (8), (9), and (10). While not necessarily the best approach for all purposes, and while other combinations of blocks are possible, the new data set created by the pooling creates a fairer basis for comparison of the ETT estimators presented in this section and the alternative estimators we discuss below.

Table 3 presents the ETT estimate for each block in the pooled data set, along with the global ETT estimates reported in the previous table. The global ETT estimates are now positive, though still smaller than the global ITT estimates. As before, this is due the influence of large blocks with negative ETTs; within blocks, as the entries in the table suggest, the ETT will have the same sign and a larger absolute value than the ITT (see equation7).[20] Again, the weighting matters greatly: we must weight the block-level treatment effects by the number of villages in each block.

[20] Again, this assumes that a greater fraction assigned-to-monitoring villages are actually monitored than assigned-to-control villages.

Table 3: Votes for Megawati: ITT and ETT, pooled "superblocks"

Block (district)	ITT	α_j, β_j	ETT
1. K. Banda Aceh, K. Makassar	24.7	$\frac{5}{47}, \frac{3}{189}$	272.8
2. Kampar and K. Pekan Baru, K. Padang	50.1	$\frac{9}{73}, \frac{1}{273}$	418.5
3. Sampang, K. Ternate	143.8	$\frac{10}{52}, \frac{1}{186}$	769.4
4. Cianjur, K. Bitung	-56.2	$\frac{5}{80}, \frac{7}{323}$	-1982.1
5. K. Palangka Raya, K. Pontianak	-209.8	$\frac{5}{15}, \frac{7}{50}$	-1,324.8
6. K. Kediri, Palembang	-75.3	$\frac{22}{35}, \frac{1}{114}$	-121.5
7. K. Mataram, Tabanan	207.0	$\frac{10}{38}, \frac{0}{102}$	786.5
8. K. Yogyakarta, K. Samarinda	23.1	$\frac{8}{28}, \frac{2}{59}$	91.8
9. Situbondo, K. Medan	212.9	$\frac{15}{62}, \frac{2}{230}$	913.0
10. K. Yogyakarta, K. Samarinda	131.9	$\frac{7}{52}, \frac{1}{167}$	1,025.6
$\text{ITT}_{\text{total votes}}$ (s.e.)	134,048.4 (153,238.6)	$\text{ETT}_{\text{total votes}}$ (s.e.)	111,835.1 (155,669.8)
$\text{ITT}_{\text{votes per village}}$ (s.e.)	61.9 (70.8)	$\text{ETT}_{\text{votes per village}}$ (s.e.)	51.7 (71.9)
$\text{ITT}_{\text{vote share}}$ (s.e.)	0.012 (0.013)	$\text{ETT}_{\text{vote share}}$ (s.e)	0.010 (0.014)

5 Instrumental-variables regression with block "fixed effects"

The ETT estimator in equation (7) is an instrumental-variables estimator. Suppose we regress votes for Megawati, by village, on an intercept and a dummy variable T_i, which equals one if the village was visited by election monitors and zero otherwise. Using a dummy variable for treatment assignment (whether a village was assigned to monitoring or not) as an instrumental variable for T_i, the instrumental-variables least squares (IVLS) estimate of the coefficient on T_i will coincide with the ETT estimator given by equation (7), within each block.[21]

However, a common approach to analyzing such experimental data using instrumental vari-

[21] See Freedman (2006) for a relevant theorem.

ables least squares (IVLS) regression is potentially misleading. In this section, we describe this alternate IVLS technique and contrast it with the global ETT estimators discussed in the previous section. Using both our original and the "pooled" data set discussed above, we show that this alternate technique produces estimates that are markedly different than the global ETT estimators discussed above.

The common approach to analyzing an experiment with block randomization is the following. The analyst supposes that the response of each village to treatment – that is, being visited by a team of election monitors – is described by the regression equation

$$M_{ij} = \alpha_j + \beta T_{ij} + \epsilon_{ij}. \tag{11}$$

Here, M_{ij} is votes for Megawati in village i in block j. On the right-hand side of the equation, β is a regression coefficient, and T_{ij} is a dummy variable that equals one if village i in block j was visited by election monitors, and zero otherwise. The intercept α_j is a "fixed effect" for block j. Under the assumptions of the model, the random error term ϵ_{ij} is independently and identically distributed (i.i.d.) across villages, with $E(\epsilon_{ij}) = 0$. Here, however, T_{ij} may be dependent on the error term, that is, endogenous: in this experiment, for instance, election observers were able to choose which villages to monitor, and their choices may be related to unmeasured factors correlated with votes for Megawati. While the Ordinary Least Squares (OLS) estimator of equation (11) will therefore be biased, random assignment of villages to monitoring may provide a valid instrumental variable.[22] Then, under the assumptions of the model, Instrumental Variables Least Squares (IVLS) regression may then provide a way to obtain consistent parameter estimates of β.

Estimation of equation (11) is misleading, however. Consider first the inclusion of the block-specific fixed effect, α_j, in the equation. In this experiment, randomization occurred within

[22]Let $Z_{ij} = 1$ if village i in region j is randomly assigned to treatment and $Z_{ij} = 0$ if the village is assigned to control. The random variable Z_{ij} is a valid instrumental variable if $\text{Cov}(Z_{ij}, T_{ij}) \neq 0$, and if it is independent of the random errors. In symbols, $Z_{ij} \perp\!\!\!\perp \epsilon_{ij}$ for all i and all j, where A $\perp\!\!\!\perp$ B means "A is independent of B."

blocks. The baseline number of votes for Megawati also varies across blocks. Analysts may reason that including a dummy variable for each block in the regression captures the different baseline votes for Megawati across blocks, so that the regression coefficient β gives the differential effect of monitoring on Megawati's village-level vote within blocks.[23]

However, as we saw above, in this experiment the blocks are different sizes. Treatment effects may be heterogenous across blocks and also related to block size. The "simple average" effect of monitoring on Megawati's vote by village, within blocks, is thus of limited interest, because this effect varies across blocks, and the blocks have different sizes. Just as the simple average of block-by-block ETT estimators is misleading for purposes of estimating the effect of treatment on the treated in the global study population (compare Tables 1 and 2), estimating equation (11) by IVLS without weighting for the different sizes of blocks will lead to misleading results.

Table 4 reports IVLS estimates of equation (11), using both the original data set, with 20 blocks (first column of the table) and the pooled data set with ten "superblocks" discussed in the previous section (third column). As the table shows, the IVLS estimates differ markedly from the global ETT estimates reported above. Most dramatically, while the ETT estimator of equation (ETTvotespervillage), using the original data set, is -1,255.2, the IVLS estimator of the β in equation (11) is 576.6. In other words, while the ETT suggests a highly negative effect of monitoring on votes for Megawati; the unweighted instrumental variables estimator, with block fixed effects, suggests a highly positive effect.

A similar if less dramatic discrepancy persists when we use the pooled data set described in the previous section. According to the global ETT estimator using the pooled data, monitoring cost Megawati an estimated 51.7 votes per village (see Table 3). On the other hand, the IVLS estimator of β in equation (11), using the pooled data, is 353.7. Even using the pooled data, then – which makes for a fairer comparison between the two estimators – the IVLS estimator is larger than the global ETT estimator by a factor of 6.8.

[23]This could be valid, if blocks were of the same size or if treatment effects were homogenous across blocks.

Again, the reason for this discrepancy is that the global ETT estimator presented in the previous section weights the block-level treatment effects by a measure of block size, that is, the number of villages in the block. Because blocks differ in size and treatment effects are heterogenous across blocks, this weighting is crucial for purposes of estimating global treatment effects for the whole study population. Simply including dummy variables for the blocks in an IVLS regression is inadequate.

Table 4: Votes for Megawati: Instrumental Variables Least Squares (IVLS) regression

	Original data		Pooled data	
	Coefficient estimates (standard errors)		Coefficient estimates (standard errors)	
Monitored (T_{ij})	576.6	460.5	353.7	280.4
	(348.1)	(302.8)	(367.5)	(307.3)
Ln Registered Voters	–	720.8	–	741.2
	–	(28.2)	–	(24.7)

Finally, another fairly common approach to analyzing these experimental data is to include covariates, as in the following regression equation:

$$M_{ij} = \alpha_j + \beta T_{ij} + X_{ij}\gamma + \epsilon_{ij}. \tag{12}$$

Here, the $1 \times p$ row vector X_{ij} includes covariates, and γ is a $p \times 1$ column vector of regression coefficients; other notation is as in equation (11). The rationale for including covariates may be to reduce the variance of the estimate of β. However, while the covariates in X_{ij} should be independent of Z_{ij}, due to the randomization, they are not necessarily independent of either T_{ij} or ϵ_{ij}. For starters, then, the non-independence of ϵ_{ij} and X_{ij} implies that the row vector $[Z_{ij}\ X_{ij}]$ will not be exogenous and thus instrumental-variables estimation of equation (12) will not be valid. For these and other reasons, adding covariates in regression analyses of experimental data may not be well-advised (see Freedman 2008).

In the second and fourth columns of Table 4, we add a scalar covariate to the previous specification; we use the natural log of the number of registered voters as the covariate.[24] The IVLS estimator of β in equation (12) is 280.4, which differs from the global ETT estimator in Table 4.2 by a factor of more than four.

6 Alternative estimators

Before turning to more foundational issues of experimental design and causal inference, we discuss a final set of estimators. Section 5 suggests that a common strategy for analyzing experiments with block randomization, using IVLS regression, can be highly misleading; the global ETT estimators provide a more sensible approach. However, both the ITT and ETT estimators discussed above may be inefficient. In this section, we discuss alternative estimators that may have lower variance.

6.1 ITT: Weighting by Registered Voters

Consider first equation (1), which is the intention-to-treat estimator for Megawati's votes by village within a block. The estimator is the sum of Megawati's votes in assigned-to-monitoring villages, divided by the number of villages assigned to monitoring, minus the sum of Megawati's votes in assigned-to-control villages, divided by the number of villages assigned to control. Villages vary greatly in size in the study population, ranging from a low of 35 registered voters to a high of 59,567.[25] Assignment of a very small or very large village to treatment can greatly vary total votes for Megawati in the treatment group, particularly when the number of villages assigned to monitoring is small. The sum of Megawati's votes in assigned-to-monitoring villages, which is a key component of equation (1), may therefore be highly variable.

[24] The rationale may be that we are adjusting for differences in village size. However, while village size is balanced across the assigned-to-monitoring and control groups, in expectation, within blocks, average village size may differ across blocks.

[25] The variation is also large within blocks: in block 9 (Kota Medan), for example, the smallest village has 30 registered voters, while the largest has 30,806.

In contrast, Megawati's vote *share* – votes for Megawati divided by the number of registered voters, by village – may be substantially more stable across villages, within a block. This fact suggests an alternative, possibly more efficient estimator for the effect of assignment to monitoring. In this section, we discuss this alternative estimator, which though biased may have lower mean squared error than the ITT estimators discussed above.

Within a block, this alternative estimator is the total number of votes for the incumbent candidate, Megawati, divided by the total number of registered voters, in villages assigned to be monitored, minus total votes for Megawati divided by total registered voters, in villages assigned to control.[26] In symbols,

$$\text{ITT}_j^A = \frac{M_j^T}{reg_i^T} - \frac{M_i^C}{reg_i^C}. \tag{13}$$

where ITT_j^A stands for the alternative ("A") intention-to-treat estimator for block J. Here, M_j^T is the number of votes for Megawati in the assigned-to-monitoring villages, in block j, and reg_i^T is the number of registered voters in the assigned-to-monitoring villages in the block; M_j^C is the number of votes for Megawati in assigned-to-control villages and reg_j^C is the number of registered voters in the assigned-to-control villages, both in block j.[27]

Notice that equation (13) gives the difference across the assigned-to-monitoring and assigned-to-control groups in the the weighted average of the vote share by village, within a block, where villages are weighted by the number of registered voters. The difference between equations (1) and (13) is that in the latter, votes for Megawati in the assigned-to-monitoring and assigned-to-control villages are divided by the number of registered voters, not the number of villages.

Now, as above, we must weight the block-by-block ITT estimators by a measure of block size. Equation (13) estimates the effect of assignment to monitoring on Megawati's vote share –

[26]Note that for both groups, the total number of registered voters divided by the total number of registered voters is the same as the weighted average of Megawati's vote share by village, where the weights are the number of registered voters.

[27]The estimand is Megawati's vote share – that is, the total number of votes for Megawati divided by the number of registered voters – if all villages in the block were assigned to be monitored, minus Megawati's vote share if all villages were assigned to control.

that is, votes over registered voters – for block j; thus, the appropriate measure of block size for purposes of weighting across blocks is the number of registered voters in the block, rather than the number of villages. Thus, here the global intention-to-treat estimator for votes for Megawati is

$$\text{ITT}^{A}_{\text{total votes}} = \sum_{j=1}^{20} reg_j \text{ITT}_j, \tag{14}$$

where ITT_j is the intention-to-treat estimator for block j, and reg_j gives the number of registered voters in block j. Note that like equation (2), equation (14) estimates the effect of assignment to treatment or control on total votes for Megawati. Other estimators are defined analogously. For instance, the global intention-to-treat estimator for the vote share is

$$\text{ITT}^{A}_{\text{vote share}} = \frac{\text{ITT}^{A}_{\text{total votes}}}{\sum_{j=1}^{J} reg_j}, \tag{15}$$

which, like equation (4), estimates the effect of assignment on Megawati's total vote share in the study population.

Table 5 reports estimated intention-to-treat effects for each of the blocks in the study, along with the number of registered voters in each block; here, we use the original rather than the pooled data set.[28] These block-by-block estimators are then weighted to arrive at the global ITT estimator given by equation (15). As shown in the final row of Table 5, the global ITT estimate is 0.7%.

Unlike the ITT estimators discussed in section 3, equation (13) is subject to ratio-estimator bias, and thus so are equations (14) and (15). Note that the numerator and denominator of equation (13) are both random: for instance, the numbers of registered voters in the treatment and control groups are random variables. Again, however, the degree of bias is likely to be small, and more

[28]Inspection of Table 5 suggests that treatment effects are related to block size. In fact, the estimated ITT effect is negatively correlated with the number of registered voters, by block, at -0.26. This is why it is important to weight the block-by-block ITT estimators by block size. Notice that the simple average of the block-level ITT effects in Table 5 is 0.012, which is non-trivially different from the valid estimated global intention-to-treat effect of 0.007 shown in the last row of the table.

Table 5: Estimated Intention-to-Treat (ITT) Effects, by Block

Block (district)	Registered Voters	ITT (increase or decrease in vote share)
1. Kota Banda Aceh	173,265	0.007
2. Kota Surabaya	2,078,486	0.009
3. Kota Mataram	241,483	0.019
4. Sampang	569,216	0.065
5. Tabanan	325,701	0.009
6. Situbondo	488,633	0.074
7. Kota Yogyakarta	327,873	0.026
8. Kota Kediri	200,137	0.059
9. Kota Medan	1,525,526	-0.008
10. Kampar and Kota Pekan Baru	740,924	0.013
11. Kota Samarinda	453,693	0.008
12. Cianjur	1,378,863	-0.013
13. Kota Palangka Raya	123,596	0.027
14. Kota Pontianak	371,780	-0.013
15. Kota Padang	525,422	-0.005
16. Palembang	906,169	-0.017
17. Kota Bitung	120,637	-0.025
18. Kota Ternate	95,771	-0.009
19. Kota Ambon	192,097	-0.050
20. Kota Makassar	812,977	0.005
$\text{ITT}_{\text{vote share}}$ (s.e.)	11,652,249	0.007 (0.0023)

appreciable for the treatment villages than the control villages. On the other hand, the estimators in equations (14) and (15) should have lower variance than their counterparts in equations (2) and (4) and in consequence may have lower mean squared error. Because the assigned-to-treatment groups are reasonably large, the jackknife will be more reliable than for the ETT estimator discussed in section 4; we estimate the standard errors using a procedure analogous to the one outlined in a note above.

Using the lower-variance estimation procedure discussed in this section, the global ITT estimate is positive and significantly different from zero at standard levels. Thus, the analysis suggests a small, positive, and statistically-significant effect of assignment to treatment or control

on the incumbent's vote share: on average, intention-to-treat by election monitoring slightly helped Megawati in the presidential runoff of 2004.

6.2 ETT: Weighting by Registered Voters

As in section 3, the ITT estimator discussed in the previous sub-section is again likely to give a conservative estimate of the effect of election monitoring, due to crossover. Across the study population, just 23.6 percent of registered voters in villages assigned to monitoring had their villages actually visited by election monitors, while 3.3 percent of registered voters in villages assigned to control had their villages monitored. In this subsection, we therefore discuss estimation of the effect of treatment on the treated – that is, the effect of treatment on villages that are monitored if assigned to monitoring and are not monitored if assigned to control – under the alternative weighting procedure discussed in this section.

Using the notation in equation (13), let ITT_j^A be the alternate ITT estimator for block j. Denote the fraction of registered voters in the assigned-to-monitoring group who reside in villages that are actually monitored as X_j^T; denote the fraction of registered voters in the assigned-to-control group who reside in villages that are (mistakenly) monitored as X_j^C. Then the alternate ETT estimator for block j is

$$\hat{\text{ETT}}_{\text{j}}^{A} = \frac{\text{ITT}_{\text{i}}}{X_j^T - X_j^C} \tag{16}$$

Table 6 reports the ETT estimator in equation (16) for each block.

Just as with the global ITT estimators, calculating the global ETT estimators requires a weighted average of the block-by-block ETT estimators; here, the weights are the numbers of registered voters in each block. For instance,

$$\text{ETT}_{\text{total votes}}^{A} = \sum_{j=1}^{20} reg_j \text{ETT}_j^A, \tag{17}$$

and

$$\text{ETT}^{A}_{\text{vote share}} = \frac{\text{ETT}^{A}_{\text{total votes}}}{\sum_{j=1}^{20} reg_j}. \tag{18}$$

The weighted average in equation (18) gives a global ETT estimator of -0.103 for the vote share (final row of Table 6).[29] For the ETT, the jackknifed estimator of variance is likely to be unreliable, due the small size of the assigned-to-monitoring group and especially the low compliance rate; we therefore use the procedure outlined in section 4, pooling within-block nominal SEs from the bivariate IVLS regression. Here, for example,

$$\widehat{\text{var}}(\text{ITT}_{\text{vote share}}) = \frac{\sum_{j=1}^{20}(reg_j)^2 \text{var}(\text{ETT}^{A}_{j})}{(\sum_{j=1}^{20} reg_j)^2}, \tag{19}$$

where $\text{var}(\text{ETT}^{A}_{j})$ is the within-block variance. Using this procedure, the estimated standard error for the global ETT estimator of 0.103 is 0.027, so the estimated effect of treatment on the treated is highly statistically significant.[30]

In sum, using the lower-variance estimator introduced in this section, the intention-to-treat effect is positive and statistically signficant, while the estimated effect of treatment on the treated is both very negative and highly significant. Why does this discrepancy occur? Again, treatment effects are heterogenous across blocks, and large blocks with sizeable ITTs also have low rates of monitoring and thus large ETTs (in absolute value). Weighting blocks and correcting for crossover involves adjustments to the data, and choices are involved in the formulation of the correction. As this example shows, different choices can give very different answers.

[29] Just as with the ITT, note that the weighting is crucial. The difference between the (correct) weighted global ETT estimator and the simple average of the block-by-block ETT estimators in Table 6 is dramatic: while the weighted ETT estimator is -0.103, the simple average of the block-by-block estimators is just -0.007.

[30] Jackknifing the variance produces a smaller estimated standard error of 0.013; again, however, the jackknife is likely to be unreliable, due to the small size of the assigned-to-treatment group and the low compliance rate.

Table 6: Estimated Effects of Treatment on the Treated (ETT), by Block

Block (district)	Registered Voters	ETT (increase or decrease in vote share)
1. Kota Banda Aceh	173,265	0.076
2. Kota Surabaya	2,078,486	0.161
3. Kota Mataram	241,483	0.040
4. Sampang	569,216	0.605
5. Tabanan	325,701	0.041
6. Situbondo	488,633	0.180
7. Kota Yogyakarta	327,873	0.132
8. Kota Kediri	200,137	0.067
9. Kota Medan	1,525,526	-0.040
10. Kampar and Kota Pekan Baru	740,924	0.118
11. Kota Samarinda	453,693	0.020
12. Cianjur	1,378,863	-1.517
13. Kota Palangka Raya	123,596	0.275
14. Kota Pontianak	371,780	-0.036
15. Kota Padang	525,422	-0.035
16. Palembang	906,169	-0.037
17. Kota Bitung	120,637	-0.088
18. Kota Ternate	95,771	-0.015
19. Kota Ambon	192,097	-0.186
20. Kota Makassar	812,977	0.095
$ETT_{\text{vote share}}$	11,652,249	-0.103
(s.e.)		(0.027)

7 The importance of experimental design

The analysis above demonstrates the sensitivity of different corrections to experimental data to the models used for adjustment. Using the number of villages to weight estimates across blocks, the global ITT for the study population gives one answer to the question, ”did (assignment to) election monitoring increase or decrease the incumbent’s vote share?” The ETT gives a very different answer to the question. Both estimators may be sensible, while the common strategy of using IVLS to estimate a regression equation with block fixed effects plainly is not. Using an alternative weighting scheme, we can achieve lower-variance ITT and ETT estimators, though at the cost of

introducing some bias.

Perhaps more than anything else, however, the discussion above suggests the importance of experimental design. In retrospect, two features of this experiment especially complicate its analysis. Because these features often seem to arise in applied work, particularly in field experiments, we discuss them in further detail in this section.

First, recall that the desire to adjust the experimental data to estimate the effect of treatment on the treated stems from the substantial crossover in this experiment. The ratio of election observers to the number of villages in the assigned-to-monitoring group, within each block, was very low, which created substantial "non-compliance" in the assigned-to-monitoring villages, as many such villages could not be visited by election observers.

An alternative to model-based adjustment is adjustment to the experimental design. In this context, that might entail reducing the number of blocks, and increasing the number of electoral teams assigned to each block, in order to boost compliance with the experimental protocol. That is, the goal is to increase the ratio of villages actually monitored to villages in the assigned-to-monitoring group. Of course, in this experiment, there may be a natural tradeoff between national representativeness and the confidence with which we can estimate treatment effects; the latter goal seems of the utmost importance to many researchers, while the former may matter most to international election observers.

A second salient feature of this experiment was that election monitors were allowed to choose the villages they visited. This may have had at least two possibly undesirable consequences. First, allowing monitors to choose villages may attenuate treatment effects. Suppose, for instance, that monitors choose villages that are easiest to reach – and that, perhaps in consequence, are less likely sites for electoral fraud. While intention-to-treat and other estimators may be unbiased (monitors would also have chosen easy-to-reach villages in the control group, had control villages been assigned to monitoring), treatment effects may be smaller due to the fact that monitors are less likely to interrupt fraud in the easy-to-reach villages where fraud tends not to occur. Second,

allowing observers latitude over the villages they monitor also heightens the possibility that observers will monitor control villages, leading to crossover from the treatment to control arm of the experiment. From a design perspective, it is therefore optimal to limit the discretion that observers have to choose the villages they monitor.

Of course, these theoretical points about design notwithstanding, there may be substantial logistical challenges involved in enforcing an experimental protocol. Partner organizations (here, the Carter Center) must weigh the value of sticking to the experimental protocol against their other concerns; together with researchers, the organization must realistically evaluate the number of units it can treat, so that the size of the assigned-to-treatment group can be chosen accordingly. Given fixed resources, there may be a tradeoff between the desire of a partner organization to expand the universe of the study (for example, by providing election monitors to a reasonably wide variety of districts) and the ability to estimate causal effects accurately. Enlisting the cooperation of partner organizations can be challenging but may the most important step, as the consequences of design choices for inferential leverage are large; these consequences should be borne in mind as analysts design and attempt to enforce experimental protocols.

8 What is the relevant parameter?

The discussion above also raises a more foundational philosophy-of-inference question: what is the relevant treatment effect? Is it the intention-to-treat parameter – here, the difference between the votes the incumbent would have obtained if all villages were assigned to monitoring and the votes she would have obtained if none were assigned to monitoring? Or is it instead the effect of treatment on the treated, that is, the effect on compliers?

In some settings, the effect of treatment on the treated may appear the more relevant parameter. After all, in many experiments there is substantial crossover, that is, some subjects assigned to treatment are subjected to the control regime, while some subjects assigned to control instead

receive the treatment. With crossover, intention-to-treat analysis usually gives a conservative estimate of the effect of treatment.

Yet intention-to-treat analysis may be the most relevant for both social-scientific and especially policy purposes. Intention-to-treat analysis, in some sense, tells us the effect of what "we" do, rather than the effect of what "they" do. It tells us, for instance, how much a given deployment of resources in a given context might be expected to affect the vote share of a political incumbent. In many settings, the effect of what "we" as social scientists – or as advisors planning an electoral monitoring mission – may be the most relevant parameter, certainly for policy purposes. Most importantly, the estimation of this parameter may involve fewer and weaker assumptions in the analysis of experimental data.

9 Conclusion

Experiments may always be analyzed according to the intention-to-treat principle, in which units randomly assigned to receive treatment are compared to those randomly assigned to control. Yet intention-to-treat analysis usually provides a conservative estimate of the effect of treatment, when compliance with an experimental protocol is low. This may lead analysts to seek to estimate the effect of treatment on the treated, that is, the effect on units that receive the treatment if assigned to treatment and otherwise are subjected to the control. Estimating the effect of treatment on the treated, however, raises issues akin those raised by observational studies, because who receives treatment – as opposed to who is randomly assigned to treatment – may be influenced by confounding factors.

The broad point we make in this article is that adjustments to experimental data involve choices, and these choices have important consequences for inferences about causal effects. Experimental design issues can substantially complicate estimation of the effect of treatment on the treated. Once we move beyond intention-to-treat analysis, which involves the purest experimental

comparison, different models may give very different results.

The experiment discussed in this paper provides a particularly intriguing example, in which there are large discrepancies between estimators of the intention-to-treat effect and the effect of treatment on the treated. These estimators can even have the opposite sign. In addition, both estimators give very different results than a common but misleading approach, which is to analyze experiments with block randomization using instrumental variables least-squares regression and block "fixed effects." The discrepancies between estimators arises for several reasons. First, there was block randomization, and blocks were of unequal sizes; next, treatment effects were heterogenous and related to block size. Finally, the monitoring rate in the treatment group was unequal across blocks, and this too was related to the size of blocks.

Analysis of this experiment illustrates a variety of other ways in which judgment enters the adjustment process. For instance, one must choose which weight to use to pool blocks for purposes of global estimators of treatment effects in the study population. The choice of weights can lead to estimators with very different properties; there can be tradeoffs between minimizing bias and variance involved in the choice of estimators.

A preferable alternative to model-based adjustment may be to alter the experimental design, where possible. For example, issues raised by crossover can be attenuated by efforts to enforce the experimental protocol. To be sure, ensuring that those assigned to treatment receive the treatment, while those assigned to control are subjected to the control, is not always feasible. Yet in field experiments like the one discussed in this paper, sensible design modifications could in principle substantially limit non-compliance. Here, for instance, one might seek to assign more electoral teams to each block (resulting perhaps in fewer monitored blocks), so that a greater proportion of villages in the assigned-to-monitoring group would be actually monitored.

The sensitivity of results to modeling assumptions raises an additional issue. Given the wide range of possible findings that result from different modeling choices in this application – in one analysis, monitoring helps Megawati a lot, while in another, monitoring hurts Megawati a

lot – one could conceive of analysts being tempted to report only those analyses that support their theoretical claims. At the least, even if multiple (and conflicting) results are reported, adjustments for multiple comparisons may be necessary if reported statistical significance tests are to be interpretable in standard ways. This point may suggest the value of specifying (and even posting on a public website, as other scholars have proposed) the hypotheses that are to be tested in any experiment, as well as the data-analytic procedures that will be used to test them. Otherwise, it is conceivable that analysis of experimental data could give rise to the kinds of specification searches that are familiar – but troubling – from modeling on observational data.

Finally, experiments like the one discussed here raise philosophy-of-inference questions. For instance, what is the most relevant parameter, intention-to-treat or the effect of treatment on the treated? In some contexts, intention-to-treat may have the greatest social-scientific as well as policy relevance. After all, given fixed resources, we might like to know what is the causal effect of what "we" as social scientists or policy advisors do. Beyond relevance, intention-to-treat analysis is the most robust way of analyzing experiments, since it leverages the purest experimental comparison.

References

[1] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-72.

[2] Freedman, David. 2000. "Notes for 215 on regression and the jackknife." Lecture notes, Department of Statistics, University of California, Berkeley. Available online at http://www.stat.berkeley.edu/ census, downloaded July 13, 2008.

[3] Freedman, David. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691-713.

[4] Freedman, David. 2008a. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40: 180193.

[5] Freedman, David. 2008b. "Notes on ratio estimators and the delta-method." Lecture notes, Department of Statistics, University of California, Berkeley. Available online at http://www.stat.berkeley.edu/ census, downloaded July 13, 2008.

[6] Freedman, David A., D.B. Petitti, and J.M. Robins. 2004. "On the Efficacy of Screening for Breast Cancer." *International Journal of Epidemiology* 33: 43-73.

[7] Freedman, David, Robert Pisani, and Roger Purves. 1998. *Statistics*. New York: W.W. Norton Company, Third Edition.

[8] Gerber, Alan, and Donald Green. 2006 "Field Experiments and Natural Experiments." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier (eds.), *Oxford Handbook of Political Methodology.* New York: Oxford University Press, pp. 357-381.

[9] Green, Donald. 2009. “Regression Adjustments to Experimental Data: Do David Freedmans Concerns Apply to Political Science?” Paper presented at the annual meetings of the Political Methodology Society, Yale University, July 2009.

[10] Hyde, Susan. 2008. “Randomizing International Election Observation: The 2004 Presidential Elections in Indonesia.” Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, Il, April 4, 2008.

[11] Imbens, Guido W. and Joshua Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62: 467-75.