The ANNALS of the American Academy of Political and Social Science

http://ann.sagepub.com/

Strategies for Dealing with the Problem of Non-overlapping Units of Assignment and Outcome Measurement in Field Experiments

Ana L. De La O and Daniel Rubenson

The ANNALS of the American Academy of Political and Social Science 2010 628: 189

DOI: 10.1177/0002716209351525

The online version of this article can be found at:

http://ann.sagepub.com/content/628/1/189

Published by:

SAGE

http://www.sagepublications.com

On behalf of:

American Academy of Political and Social Science

Strategies for Dealing with the Problem of Non-overlapping Units of Assignment and Outcome Measurement in Field Experiments

By
ANA L. DE LA O
and
DANIEL RUBENSON

Researchers conducting field experiments are sometimes faced with the challenge of analyzing field experiment results when the unit of assignment does not coincide with the unit of outcome measurement. For example, in electoral research, election results may be reported at a level of geography defined by electoral law, while the assignment of treatment can be made only at a level of geography different from this. Using examples from field experiments conducted in Canada and Mexico, we describe this problem and its consequences for analysis and interpretation of field experiment data and results. We also offer a number of practical solutions analysts can employ when faced with non-overlapping units of assignment and outcome measure in field experiments.

Keywords: causal inference; field experiments; treatment density; failure-to-treat; SUTVA; non-overlapping units

In the study of electoral politics, as well as other domains, researchers will sometimes be faced with institutional and political restrictions to the randomization of the unit of outcome measurement. In such cases, researchers can design a field experiment where the unit of assignment to treatment does not coincide with the unit of outcome measurement. To address the problem on non-overlapping units, at first

Ana L. De La O is an assistant professor of political science at Yale University. She is also affiliated with the Institution for Social and Policy Studies and the MacMillan Center for International and Area Studies at Yale. Her current research interests include politics of poverty alleviation, public goods provision, and clientelism.

Daniel Rubenson is an assistant professor of political science at Ryerson University, Toronto. His research is mainly concerned with questions about political campaigns and communication, leadership, elections, and participation. He mainly employs field experiments in studying these issues and has carried out several experiments in the context of various campaigns in Canada, including referendums and a party leadership convention. His work has appeared in Party Politics, Electoral Studies, *and* Acta Politica.

DOI: 10.1177/0002716209351525

glance, an appealing possibility is to conduct a survey in treated and control units to measure the outcome. However, not only is a survey expensive, but self-reported electoral behavior is problematic at best and misleading at worst (Bernstein, Chadha, and Montjoy 2001; Karp and Brockingtona 2005).

Field experimentalists have the alternative of using official reported electoral results, but in doing so they will have to face at least three challenges. First, researchers will have to overlay the units of assignment to treatment and control into the unit of outcome measure. Depending on the application, this process could be either straightforward or taxing. If the units of assignment to treatment and control overlap perfectly to the unit of outcome measure, the control group continues to be a valid counterfactual for the treatment group and no special analysis is needed. If, however, the unit of outcome measure and the unit of assignment do not overlap perfectly, the aggregation into the unit of outcome measure may bring in units that were originally excluded from the experimental sample. Once aggregated, treatment status must be accompanied by a fractional response variable that captures the density of the treatment. In such cases, even if the original randomization was successful, the density of the treatment is not random and could be correlated with observed and unobserved covariates in a way that would render causality implausible. Finally, the problem of non-overlapping units introduces the possibility of violating the assumption known as the stable unit treatment value assumption (SUTVA) (Rubin 1980, 1986), if spillover effects are present.

In this article, we describe the non-overlapping problem using two examples from recent research in Mexico and Canada: De La O (2008) and Loewen and Rubenson (2008). Fortunately, the statistical and evaluation literature offers researchers several practical solutions to these problems. We present these solutions in an intuitive way. For the formal solutions, or the applications of these solutions to the Mexican and Canadian cases, please refer to the original articles. The article is organized in four sections, each presenting a part of the problem and its solutions. We proceed from the easier problem to the more challenging one. In the next section, we introduce the overall problem. We then present the challenge of estimating average treatment effects when units are treated with different dosages, describe the consequences of violating SUTVA, and then offer our conclusion.

The Problem

As the articles in this volume—and the growing literature on experimentation in political science more generally—attest, field experiments come with many advantages. Their most important feature is the random assignment of a treatment of interest to a well-defined population, followed by a statistical analysis of the effects of the treatment (Green and Gerber 2004). Random allocation not only provides a justification for causal inference, but also ensures that treatment and control units have similar observed and unobserved baseline characteristics.

But what happens when treatment is assigned to a population that is less well defined or, indeed, different from that on which the outcome is measured? For example, De La O (2008) analyzed the effect of a Mexican anti-poverty program on turnout and vote shares, taking advantage of a randomized experiment conducted by the program's operators in the early phases of the program. The experiment was designed to evaluate the effects of the program on children's schooling and health. Randomization was done at the village level, but electoral results in Mexico are reported at the polling precinct level. Similarly, Loewen and Rubenson (2008) analyzed the effects of a direct mail campaign on referendum results in the Canadian province of Ontario, where a flyer was delivered via mail. Again, election and referendum results are officially reported at the polling precinct level. However, using the polling precincts as the unit of assignment to treatment was not possible for two reasons: first, Elections Ontario (the authority administering elections in the province) determines precinct boundaries very close to Election Day, making it logistically impossible to plan ahead. At the time the field experiment was designed, the boundaries of the polling precincts were unknown. Second, because the flyers were to be delivered by mail, the researchers were limited by Canada Post delivery standards. Randomization was done at a level of geography called a "letter carrier walk" (LCW), the route postal carriers take when they deliver the mail, and LCW boundaries do not correspond to polling precinct boundaries.

In these examples, as in many other contexts of interest to social scientists, institutional, political, and logistical matters render randomization at the unit of outcome impossible. For the Mexican anti-poverty program, for instance, using the polling precincts as the unit of randomization would have fueled enough political opposition to prevent the program's scaling-up. In the Canadian example, only an institutional reform would allow for a design that allocates treatment at the polling precinct level.

In such cases, the random allocation of treatment at a different unit of analysis still provides a formal justification for causal inference, impartiality in the allocation of treatment, and transparency in such allocation. Yet, once villages or LCWs are aggregated to the polling precinct unit, there is a possibility that the baseline characteristics of polling precincts including treatment and control villages (or LCWs) are unbalanced. To improve balance, researchers can make use of widely known approaches to covariate adjustment. For example, one can include relevant covariates as control variables in the regression estimating the average treatment effect (Imbens 2004). Two-stage techniques, where the first stage estimates the propensity score—the probability of treatment—conditional on some covariates, is another option (Rosenbaum and Rubin 1983; Imbens 2000; Hahn 1998; Rosenbaum 1987; Hastings et al. 2007; Wooldridge 2004, 1999; Hirano, Imbens, and Ridder 2003; Imbens 2004). Aside from regression methods, an increasingly popular method is matching on covariates or matching on the propensity score

(Rosenbaum and Rubin 1983; Ho et al. 2007; Diamond and Sekhon 2008; Sekhon 2004; Brunell and DiNardo 2004).

Covariate adjustment of the kind described above essentially improves the statistical precision of the estimates. However, with or without covariates as controls, the treatment effect estimates that result from the analysis of randomized data are unbiased, especially given a large enough sample size. For ease of presentation, in the remainder of the article, we assume an infinite sample size so that we can focus on the inference issues specific to the aggregation problem we have outlined above.

Average Treatment Effects When Dosage Varies

In most cases with non-overlapping units, the unit of outcome tends to be larger than the unit of randomized assignment. This has consequences for analysis and interpretation of field experiment data and results because, after aggregation, the density of treatment—or *dosage*—varies across units. Consider the example in De La O (2008) where neither villages nor polling precincts have fixed population sizes. Among the experimental sample, villages are smaller than polling precincts. Estimating the treatment effect using a dummy variable taking the value of one for the polling precincts in the treatment group could be misleading, because even if treatment is random, the density of treatment may no longer be independent of background characteristics. It is relatively easy to identify ways in which this could be a problem. For example, at the precinct level, the fraction of the polling precinct treated depends on its population. The more populous the precinct, the smaller the fraction treated. Population, however, could also correlate with the level of urbanization of the polling precinct. In turn, urbanization could correlate with support for the opposition parties. In the case of Loewen and Rubenson (2008), the problem manifests itself in a similar way. More densely populated areas are more likely to have a larger number of LCWs—the unit of assignment to treatment—and therefore, these areas are perhaps more likely to receive a higher dosage of treatment. Again, population density might correlate with vote choice on the referendum question. If this logic is correct, the density of treatment is no longer independent from the error term in the equation explaining the outcome as a function of the intervention.

To shed more light onto this problem, suppose for simplicity that a researcher wants to find the average treatment effect, t, on vote share y. Consider the hypothetical example depicted in Table 1, where ten villages are randomly assigned to receive the treatment, with villages fitting into precincts. As shown in the third column, one set of villages corresponds 1:1 into precincts. A second set corresponds 2:1. Each village has a 50 percent chance of being assigned to the treatment. In the first set of precincts, this means that 50 percent are in the treatment group and 50 percent are in the control group. In the second set, 25 percent are in control, 50 percent are in half-treatment, and 25 percent are in treatment. What are the consequences of this?

TABLE 1
A Hypothetical Example of Non-overlapping Units

			Potential Outcomes		
Villages	Polling Precinct	Strata	Under the Control State	Under the Treatment State	Under the Half-Treatment State
1	A	1	1	4	.
2	B	1	1	4	.
3	C	2	2.5	10	5
4					
5	D	2	2.5	10	5
6					
7	E	2	2.5	10	5
8					
9	F	2	2.5	10	5
10					

Villages	Polling Precinct	Strata	Treatment	Polling Precinct's Treatment	Observed Outcomes
1	A	1	1	Treatment	4
2	B	1	0	Control	1
3	C	2	1	Treatment	10
4			1		
5	D	2	1	Half-treatment	5
6			0		
7	E	2	0	Half-treatment	5
8			1		
9	F	2	0	Control	2.5
10			0		
			Pr $[t = 0] = .5$	Pr $[t = 1] = .5$	

First, note that polling precincts in the two sets have different probabilities of treatment. Furthermore, as shown in the fourth column of the first panel, the potential outcome under the control state is 1 for some precincts and 2.5 for others. This reveals that precincts are unlike each other, yet the ways in which they differ are captured perfectly by the set to which precincts belong. By taking into account the stratification of the data, researchers can avoid a correlation between polling precinct size (and all that goes with it as discussed above) and treatment, which could produce bias.

Second, as shown in the second panel of Table 1, for precincts in the second set, two alternative treatments are available: treatment and half-treatment.

For this example then, one can estimate the treatment effect as the weighted average of observed outcomes for those in the treatment and control groups in the two sets as follows:

$$(4-1)(\frac{1}{2}) + (10-2.5)(\frac{1}{2}) = 5.25.$$

Thus, for this example, where precincts are stratified, one can obtain unbiased estimates of the treatment effect by appropriately weighting estimates for the two strata. Alternatively, the effect of the half-treatment intervention can be estimated as

$$(\frac{5+5}{2}) - 2.5 = 2.5.$$

Non-overlapping Units as a Failure-to-Treat Problem

The imperfect overlap of villages to polling precincts could also be framed as a failure-to-treat problem. Consider again the hypothetical example in the previous section. Among the group of precincts, we have precincts for which the treatment is unavailable, precincts for which the treatment was administered to one village, and precincts for which treatment was administered in both villages. Let the expected vote outcomes in the three groups be V_c, V_h, and V_t, respectively. Each precinct has three components: the first village, the second village, and an extraneous component. The expected vote outcome in the control group can be written as a weighted average of these three components, where the weights are w_1, w_2, and $(1 - w_1 - w_2)$ and the village vote shares are the s_j:

$$V_c = w_1 * s_1 + w_2 * s_2 + (1 - w_1 - w_2) * s_3. \tag{1}$$

The treatment group vote is expected to be

$$V_t = w_1(s_1 + t) + w_2(s_2 + t) + (1 - w_1 - w_2) * s_3, \tag{2}$$

where t is the treatment effect. The half-treatment group vote is expected to be

$$V_h = w_1(s_1 + t) + w_2(s_2) + (1 - w_1 - w_2) * s_3, \tag{3}$$

where the first village is assumed to be the one that is randomly selected for treatment.

With this formalization of the problem, it is now straightforward to solve for t in order to generate a consistent estimator of the treatment effect. For the treated units, we solve for t as follows:

$$V_t - V_c = w_1 * t + w_2 * t, \tag{4}$$

and

$$t = \frac{V_t - V_c}{w^1 + w^2}. \tag{5}$$

Equations (6) and (7) solve for t in the case of the half-treated units:

$$V_h - V_c = w_1 * t, \tag{6}$$

and

$$t = \frac{V_h - V_c}{w^1}. \tag{7}$$

In this setup, we assume that SUTVA holds. In the following section, we discuss the consequences of dropping this assumption.

Non-overlapping Units as a SUTVA Problem

The variation in the density of treatments inevitably raises the following question: is the effect driven by treated villages within treated precincts, untreated villages within treated precincts, or a mix of both? This question is related to the stable unit treatment value assumption in the counterfactual (Rubin 1980, 1986). This assumption "requires that the potential outcomes of individuals be unaffected by potential changes in the treatment exposure of other individuals" (Morgan and Winship 2007, 37). When SUTVA is violated, the possibility of spillovers erodes the transparency of the counterfactual setup in an experiment.

Consider the Mexican anti-poverty program example where a positive treatment effect of the intervention could mean that the program is mobilizing its recipients. A positive treatment effect, however, is also consistent with a story that combines the mobilization of recipients and the demobilization of non-recipients living within treated precincts. In this case, the intervention's indirect negative effect on non-recipients would produce a diluted effect of the program.

This type of treatment effect dilution is only one example in which the violation of SUTVA can bias the estimates of the treatment effect. In other applications, spillovers could have the opposite effect. Taking the Canadian example, if spillovers are present, a political party may need to distribute flyers to only 75 percent of an LCW in order to publicize its campaign, provided that the other 25 percent learn about the campaign from their neighbors. Following this example, if spillovers are contained within polling precincts, then the spillovers will lead to an overestimation of the treatment effect. Similarly, spillovers can magnify the treatment effect in such a way that the greater the number of units treated, the more effective the intervention is.

Consider again the example above where one portion of the precincts maps 1:1 and another portion maps 2:1 to villages. Vote share in precincts in the control group is given by equation (1), with w_j and s_j described as before. If SUTVA holds, the treatment effect for the treated and half-treated units is given by equations (5) and (7). Yet, the possibility that the treatment administered to one village in the precinct will have effects on the other village remains as a concern. We will call this type of within-precinct spillover type 1. Similarly, it is problematic that the treatment administered to one precinct may have spillovers on a precinct in the control group. We will call this type of across-precinct spillover type 2.

If type 1 spillovers are present, then the treatment group vote can be expressed as

$$V_t' = w_1(s_1 + t) + w_2 * (s_2 + t) + (1 - w_1 - w_2) * (s_3 + spillover). \tag{8}$$

Subtracting equation (1) from equation (8), we get that treatment is now

$$V_t' - V_c = w_1 * t + w_2 * t + (1 - w_1 - w_2) * (spillover), \tag{9}$$

$$t = \frac{V_t' - V_c - (1 - w_1 - w_2) * (spillover)}{w^1 + w^2}. \tag{10}$$

Comparing the treatment effect from equation (10) to the treatment effect obtained in equation (5), we see that spillovers introduce a bias of $-(1 - w_1 - w_2) *$ (*spillover*). This bias is increasing in the size of the extraneous component in the precinct.

The half-treatment group vote can be expressed as:

$$V_h' = w_1 * (s_1 + t) + w_2 * (s_2 + halfspillover) + (1 - w_1 - w_2) * (s_3 + halfspillover), \tag{11}$$

where s_1 is the treated village. Now treatment effects are

$$V_h' - V_c = w_1 * t + w_2 * (halfspillover) + (1 - w_1 - w_2) * (halfspillover), \tag{12}$$

and

$$t = \frac{V_h' - V_c - (1 - w_1 - w_2) * (halfspillover) - w_2 * (halfspillover)}{w^1}. \tag{13}$$

Under this condition the spillovers introduce a bias of $-(1 - w_1 - w_2) *$ (*halfspillover*) $- w_2 *$ (*halfspillover*), which is increasing in the size of the second village and the extraneous component in the precinct.

To illustrate this point, consider the example in Table 2, where a case of spillover type 1 is presented. While the effect of treating a precinct remains unchanged, the effect of treating half of a precinct is now

$$\left(\frac{3+3}{2}\right) - 2.5 = .5.$$

TABLE 2
AN EXAMPLE OF SPILLOVER TYPE 1

				Observed Outcomes	
Villages	Polling Precinct	t	Polling Precinct's Treatment	No Spillover	Spillover
1	A	1	Treated	4	4
2	B	0	Control	1	1
3	C	1	Treated	10	10
4		1			
5	D	1	Half-treated	5	3
6		0			
7	E	0	Half-treated	5	3
8		1			
9	F	0	Control	2.5	2.5
10		0			

Thus, the effect of the half-treatment is now diluted. Now, consider a case where a spillover type 2 is present. The vote share in the control precincts can be written as

$$V'_c = w_1 * (s_1 + spillover) + w_2 * s_2 + (1 - w_1 - w_2) * s_3. \tag{14}$$

Here, the first village receives the spillover effect of a village treated in a different precinct. By substituting V_c for V'_c in equation (5), we now get a diluted treatment effect since the spillover makes V'_c larger than V_c. Consider the example in Table 3, where a spillover type 2 is present. In this example, the effect of the treatment is

$$(4 - 2)(\tfrac{1}{2}) + (10 - 3.5)(\tfrac{1}{2}) = 4.25.$$

And the effect of the half-treatment is

$$(\frac{5+5}{2}) - 3.5 = 1.5.$$

As this section shows, the presence of spillovers makes the identification of the treatment parameter impossible without some additional strong assumptions. As Morgan and Winship point out, for many applications, SUTVA is a restrictive assumption (2007, 37-39). Consider again the poverty relief program example. The suitability of SUTVA may depend on the scale of the program. For a small program situated in a large country, the vote share of non-recipients may be entirely independent from the existence of the program. Yet, a sizable program in a small country may induce changes both in recipients' and non-recipients' electoral behavior.

TABLE 3
An Example of Spillover Type 2

Villages	Polling Precinct	t	Polling Precinct's Treatment	Observed Outcomes: No Spillover	Observed Outcomes: Spillover
1	A	1	Treated	4	4
2	B	0	Control	1	2
3	C	1	Treated	10	10
4		1			
5	D	1	Half-treated	5	5
6		0			
7	E	0	Half-treated	5	5
8		1			
9	F	0	Control	2.5	3.5
10		0			

If there are doubts as to whether SUTVA will hold in a particular intervention, one possibility researchers may find attractive is to conduct two surveys, one prior to and one after the intervention, to detect changes in outcomes for the control group compared to the treatment group. Another possibility researchers may find useful is to incorporate the spillover effects into the design of the experiment. In fact, researchers may be interested in estimating these effects explicitly. For instance, in the Canadian example, researchers may be interested in estimating the spillover effects created by varying the density of treatments from, say, 20 to 80 percent of a LCW. This estimation of spillovers is, however, a challenging and expensive enterprise as the number of observations needed to get sufficient power to detect one-fifth and fourth-fifths of the treatment effect may be prohibitive. Finally, an alternative to varying the density of treatments is to design the intervention to capture its spatial diffusion.

Conclusion

In the study of electoral politics, researchers will sometimes be faced with institutional and political restrictions to the implementation of field experiments when the unit of assignment does not coincide with the unit of outcome measurement. For example, in the study of the effects of government-funded poverty relief programs, which by design are non-political, or the study of elections in places where it is the institutional tradition to determine the boundaries of the polling precincts quite late, or where the delivery of treatment can be made only at a level of geography different from polling precincts, randomization at the unit where electoral results are reported is impossible. In this article, we present several solutions to deal with unbalance of baseline covariates, to estimate treatment effects when there is variation in treatment density, and to deal with violations of SUTVA. While the problems associated with non-overlapping units of treatment and outcome are not trivial, they ought not to be seen as impossible to overcome. With close attention paid to these issues and careful analysis, we believe that researchers ought to still pursue field experimental designs in these situations.

References

Bernstein, Robert, Anita Chadha, and Robert Montjoy. 2001. Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly* 65 (1): 22-44.

Brunell, Thomas L., and John DiNardo. 2004. A propensity score reweighting approach to estimating the partisan effects of full turnout in American presidential elections. *Political Analysis* 12 (1): 28-45.

De La O, Ana. 2008. Do poverty relief programs affect electoral behavior? Evidence from a randomized experiment in Mexico. Unpublished manuscript, Yale University, New Haven, CT.

Diamond, Alexis, and Jasjeet S. Sekhon. 2008. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Unpublished manuscript, University of California, Berkeley.

Green, Donald P., and Alan S. Gerber. 2004. *Get out the vote*. Washington, DC: Brookings Institution.

Hahn, Jinyong. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66 (2): 315-31.

Hastings, Justine S., Thomas J. Kane, Douglas O. Staiger, and Jeffrey M. Weinstein. 2007. The effect of randomized school admissions on voter participation. *Journal of Public Economics* 91:915-37.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161-89.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199-236.

Imbens, Guido W. 2000. The role of propensity score in estimating dose-response functions. *Biometrika* 87 (3): 706-10.

Imbens, Guido W. 2004. Nonparametric estimation of average treatment effects under exogenity: A review. *Review of Economics and Statistics* 86 (1): 4-29.

Karp, Jeffrey A., and David Brockingtona. 2005. Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries. *Journal of Politics* 67:825-40.

Loewen, Peter John, and Daniel Rubenson. 2008. Both sides now: A field experiment with competing messages. Unpublished manuscript, Ryerson University, Toronto, Canada.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

Rosenbaum, Paul R. 1987. Model-based direct adjustment. *Journal of the American Statistical Association* 82:387-94.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41-55.

Rubin, Donald B. 1980. Comment on "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association* 75:591-93.

Rubin, Donald B. 1986. Which ifs have causal answers? *Journal of the American Statistical Association* 81:961-62.

Sekhon, Jasjeet S. 2004. The varying role of voter information across democratic societies. Unpublished manuscript, University of California, Berkeley.

Wooldridge, Jeffrey M. 1999. Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* 67:1385-1406.

Wooldridge, Jeffrey M. 2004. Estimating average partial effects under conditional moment independence assumptions. Unpublished manuscript, Michigan State University, East Lansing.