Technical Note on the Conditions Under Which It is Efficient to Discard Observations Assigned to Multiple Treatments in an Experiment Using a Factorial Design

Alan S. Gerber
Donald P. Green
Yale University

January 4, 2003

Much of Imai's (2003) paper takes umbrage at the New Haven study's (Gerber and Green 2000) use of factorial design. Gerber and Green (2004, Table A1) show that none of the substantive conclusions of the New Haven study hinge on the inclusion or exclusion of those assigned to multiple treatments. Nevertheless, the issue of whether to use factorial designs is an important one, and it is important to dispel the misconceptions that Imai generates concerning the design and analysis of factorial experiments.

Experimental researchers routinely use factorial designs in order to gauge interactions between factors. The use of factorial design is particularly common among researchers launching new research agendas. As Mead (1988, pp.584-5) explains in *The Design of Experiments: Statistical Principles for Practical Applications:*

> A major question will usually be how many aspects of variation, or factors, to consider simultaneously in an experiment. This will depend on the stage of the experiment within the research programme. At the begining of the programme, it is important to include many factors, because failure to investigate the interaction of factors at an early stage can easily lead to wasting resources through pursuing too narrow a line of research before eventually checking on the effects of other

factors. It is difficult to conceive of an investigation with too many factors at an early stage.

This research trajectory is typically combined with a hypothesis-testing approach in which analysts look for evidence of interactions and, failing to find them, focus on main effects. This approach figures prominently in virtually all textbooks on experimental design and data analysis (Box, Hunter, and Hunter 1978; Cochrane and Cox 1957; Mead 1988; Montgomery 2003), and it is the framework within which the Gerber and Green experiment was designed and analyzed.

The variegated factorial design of the study enabled us to investigate a wide variety of potential interactions suggested in previous work that also used factorial design (Eldersveld 1956; Miller et al. 1981). Factorial design also enabled us to approximate real world variation in background campaign conditions. When examining the effects of phone calls for the population that also receives three mailings, one seeks to learn about the effects of phone calls in environments where other campaigns have sent out mailings.

Breaking with conventional thinking about factorial design, Imai derides the use of multiple treatments as "incorrect" and "inefficient." He declares on p.9: "In principle, it is advisable to minimize the number of treatments in field experiments." He cannot imagine why we would waste observations on multiple treatments, apparently forgetting that the treatment and interaction effects reported in the studies that he cites with approval (Eldersveld 1956; Miller et al. 1981) were often very large. It is important to note that our experiment had the power to detect the massive interactions that Miller et al. report (1981, p.450). It turned out, however, that our experiment revealed no significant interactions (p.660), and so we focused our write-up on main effects, but the absence of robust interactions is a finding, not a design failure.

After donouncing factorial design, Imai's proposes to discard all the observations in the New Haven study that were assigned to more than one type of treatment. Imai provides no analytics to bolster the efficiency of this statistical approach, nor does he cite

any authorities on this point. His argument rests entirely on the substantive assertion that "treatment effects are likely to be smaller for those who have already received other treatments" (footnote 6). This leaves the reader to fill in the missing statistical argument raised by the question: Under what circumstances is it more efficient to discard or retain observations assigned to multiple treatments? Because this question has broad implications for experimental design and analysis, we welcome the opportunity to discuss the answer, which follows directly from standard econometric treatments of when to include or exclude potential regressors. Here, we show the implications of this analytic result for the New Haven experiment with special reference to the effects of phone calls.

Let $\hat{b}_1$ be the instrumental variables estimator of the phone treatment effect, β, based solely on what Imai terms the "correct" groups (i.e., the control group that receives no treatment whatsoever and the treatment group that receives only phone calls). This estimator is what Imai presents in Tables 5. Let $\hat{b}_2$ denote the instrumental variables estimator of the phone treatment effect based on a comparison of "incorrect" groups (e.g., the control group that received mail and the treatment group that received both mail and phone calls). We will use γ to denote the bias associated with $\hat{b}_2$, allowing for the possibility that "treatment effects are likely to be small for those who have already received other treatments" (Imai 2003, footnote 6). Call $\hat{b}_T$ the instrumental variables estimator of the phone treatment effect, b_1, based on both types of data; this estimator is what Gerber and Green (2000) use. It is helpful to express $\hat{b}_T$ as a weighted average of $\hat{b}_1$ and $\hat{b}_2$. Using σ_1^2 to denote the sampling variance of $\hat{b}_1$, and σ_2^2 to denote the sampling variance of $\hat{b}_2$, we express $\hat{b}_T$ as follows:

$$\hat{b}_T = \hat{b}_1 \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \hat{b}_2 \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

Having defined the competing estimators, we now compare their properties. The estimator $\hat{b}_1$ is an unbiased estimator of β with variance σ_1^2. Thus, its mean squared error (MSE) is also σ_1^2. By comparison, $\hat{b}_T$ has an expectation of $\beta + \gamma \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$ and a variance of $\frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$. The MSE of $\hat{b}_T$ is therefore

$$MSE\left(\hat{b}_T\right) = \gamma^2 \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^2 + \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

A comparison of $MSE\left(\hat{b}_1\right)$ and $MSE\left(\hat{b}_T\right)$ shows that the using all of the data is more efficient than using only the "correct" groups when

$$|\gamma| < \sqrt{\sigma_1^2 + \sigma_2^2}.$$

The right hand side of this equation is recognizable as the standard error of the difference between the estimated treatment effects in the correct and incorrect groups. This formula thus provides a simple decision rule: *Discard the "incorrect" group if one suspects that the treatment effect in the incorrect group lies more than one standard error away from the treatment effect in the correct group*.

Is this decision rule satisfied in the New Haven study? From Imai's analysis of the original New Haven data (Table 5), we learn that $\hat{\sigma}_1$ is 6.6, which means that the bias due to the use of incorrect groups must be *at least* 6.6 percentage-points in order to warrant discarding the observations assigned to multiple treatments. If the true parameter were 5, as Imai contends in his abstract, the ex ante claim that the use of multiple groups leads to a downward bias of more than 6.6 means that phone calls diminish turnout among those who receive other treatments. As it turns out, $\hat{b}_1 < \hat{b}_2$ in the original data,

which is the opposite of what Imai's diminishing returns argument supposes. Using the corrected data and robust standard errors, we noted earlier that $\hat{b}_1$ is -0.9 and $\hat{b}_t$ is –1.6, which implies that $\hat{b}_2$ equals –2.1 percentage-points and the bias (γ) is –1.2. Since $\hat{\sigma}_1$ is 6.0 and $\hat{\sigma}_2$ is 2.6, one would not discard the observations assigned to multiple treatment groups unless the bias were greater than 6.5, and empirically the absolute bias (1.2) is nowhere near this value. Thus, a means-squared error analysis demonstrates that discarding the "incorrect" groups in the New Haven study cannot be justified on grounds of efficiency.

One could argue that Imai acted in accordance with his own very strong priors about the bias associated with multiple treatments, even if his ideas about efficiency were not well thought out. But a closer look at his views on the subject of bias reveals them to be inconsistent as well. Imai is dead set against factorial designs on pages 9-10, but by page 12 he has changed his mind and applauds his discovery that

> sending a postcard three times is much more effective than mailing it once or twice. This provides evidence against the assumption of Gerber and Green (2000) that the effect of mail canvassing is linear in the number of postcards sent.

Imai's conviction about diminishing returns has now reversed itself. Leaving aside the fact that this alleged departure from linearity is not statistically significant, the irony of Imai's comment is that he would not have had the opportunity to make this observation had we obeyed his principle of minimizing the number of treatment groups.

In sum, Imai's statistical practice of discarding observations assigned to multiple treatments cannot be justified empirically, and the theoretical justification for doing so rests on mercurial assumptions. Nor do we endorse Imai's stricture (p.10) that "In principle, it is advisable to minimize the number of treatments in field experiments." This strikes us as shortsighted advice. The fact that so many of the interactions in the New Haven study proved to be inconsequential is a research finding made possible by the

exploration of many factors. To dub the use of factorial design as "inefficient" fails to appreciate the aims of the study.

References

Box, George E.P., William G. Hunter, and J. Stuart. Hunter. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: John Wiley & Sons.

Cochrane, William G., and Gertrude M. Cox. 1957. *Experimental Designs*. New York: John Wiley & Sons.

Eldersveld, Samuel J. 1956. Experimental Propaganda Techniques and Voting Behavior. *American Political Science Review* 50(March): 154-65.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Personal Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review, 94* (3): 653-64.

Gerber, Alan S. and Donald P. Green. 2004. Brief Paid Phone Calls Are Ineffective: Why Experiments Succeed and Matching Fails To Produce Accurate Estimates. Unpublished manuscript, Institution for Social and Policy Studies, Yale University.

Imai, Kosuke. 2003. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review,* final manuscript submission, August 19th.

Mead, R.. 1988. *The Design of Experiments: Statistical Principles for Practical Application*. New York: Cambridge University Press.

Miller, Roy E., David A. Bositis, and Denise L. Baer. 1981. Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman. *International Political Science Review* 2(4): 445-60.