NBER WORKING PAPER SERIES

SO YOU WANT TO RUN AN EXPERIMENT, NOW WHAT? SOME SIMPLE RULES OF THUMB FOR OPTIMAL EXPERIMENTAL DESIGN

John A. List
Sally Sadoff
Mathis Wagner

Working Paper 15701
http://www.nber.org/papers/w15701

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2010

We thank the Editor, two anonymous referees, Glenn Harrison, Emily Oster, Stephen Raudenbush, Azeem Shaikh, and seminar participants at the University of Chicago for useful comments. Clifford Clive provided research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design

John A. List, Sally Sadoff, and Mathis Wagner

NBER Working Paper No. 15701
January 2010
JEL No. C9,C91,C92,C93

ABSTRACT

Experimental economics represents a strong growth industry. In the past several decades the method has expanded beyond intellectual curiosity, now meriting consideration alongside the other more traditional empirical approaches used in economics. Accompanying this growth is an influx of new experimenters who are in need of straightforward direction to make their designs more powerful. This study provides several simple rules of thumb that researchers can apply to improve the efficiency of their experimental designs. We buttress these points by including empirical examples from the literature.

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

Sally Sadoff
Department of Economics
University of Chicago
sadoff@uchicago.edu

Mathis Wagner
Collegio Carlo Alberto
Via Real Collegio 30
10024 Moncalieri (TO)
Italy
mathis.wagner@carloalberto.org

1 Introduction

Ever since economists became engaged in the data business, they have grappled with how to construct the proper counterfactual. The concept of identifying a treatment effect is simple enough conceptually, but in practice a major problem is one of a missing counterfactual - person i is not observed in more than one state simultaneously. Within economics, measurement approaches can be divided into two main categories: estimation of models that make use of naturally-occurring data and approaches wherein the analyst herself governs the data generation process. A handful of popular empirical approaches are typically used when the analyst is dealing with naturally-occurring data, but the literature is replete with criticisms of their identifying assumptions, many times based on restrictiveness or implausibility (see Blundell and Costas-Dias, 2002, for a useful review).

In those cases where the analyst generates her own data, such as within the area of experimental economics, identification assumptions are much less severe. To obtain the effect of treatment in the particular domain of study the only major assumption necessary is appropriate randomization (with appropriate sample sizes). In this manner, when running an experiment the analyst is using randomization as an instrumental variable (see List, 2006). But, with the chore of data generation comes other, less discussed, obligations of the researcher. In this study, we consider one such feature more carefully: the optimal number and arrangement of subjects into experimental cells.

A casual perusal of the literature presents a striking consistency concerning sample sizes and their arrangement: most studies uniformly distribute at least 30 subjects into each cell. This approach holds whether the analyst is making use of a purely dichotomous treatment (i.e., pill or no pill) as well as when the analyst is exploring levels of treatment (i.e., various dosage levels). Discussion of whether such a sample arrangement is efficient is more mature in other literatures, but has not been properly vetted in the experimental economics community. Our paper attempts to fill this gap. In doing so, we do not claim originality in any of the derivations, rather this study should be viewed as a compilation of insights from other literatures that might help experimenters in economics and related fields design more efficient experiments.

Our study begins with a discussion of popular randomization techniques. We discuss the virtues of complete randomization, block designs, and factorial designs.[1] After these randomization preliminaries, we move to a discussion of the power of the experimental design. We provide simple formulas with which to compute required sample sizes under three major classes of assumptions: (1) a dichotomous treatment with potentially heterogeneous treatment effects (for continuous and binomial outcomes) (2) a dichotomous treatment in a cluster design, and (3) a continuous treatment with homogeneous treatment effects. We elaborate on these simple formulas in cases where the cost of sampling subjects differs across treatment and control and where there is a fixed cost of sampling from a new cluster.

Several simple rules of thumb fall out of the discussion. The overarching idea revolves around first implementing an experimental design that maximizes the variance of the treatment variable, and second adjusting the samples to account for variance heterogene-

[1] Fisher (1935) and Cox and Cochrane (1950) provide seminal discussions of experimental design.

ity, if necessary. In the case of a simple comparison between a single treatment and a control group, one first insight is that with a continuous outcome measure, under the null hypothesis of no treatment effect, one should only allocate subjects equally across treatment and control if the sample variances of the outcome means are expected to be equal in the treatment and control groups (i.e., in those cases when there are homogeneous treatment effects). The optimal sample arrangement becomes more lopsided as the sample variances of outcomes across treatment and control become more disparate; or likewise, the treatment effect becomes more heterogeneous. A simple rule of thumb to maximize power given a fixed experimental budget naturally follows: the ratio of the sample sizes is equal to the ratio of the standard deviations of outcomes.

In cases when the outcome variable is dichotomous, under the null hypothesis of no treatment effect (i.e. $p_1 = p_0$, where the subscripts index treatment and control respectively), one should always allocate subjects equally across treatments. This follows from the close connection between mean and variance. Yet, if the null is of the form $p_1 = kp_0$, where $k > 0$, then the sample size arrangement is dictated by k in the same manner as in the continuous case. If the cost of sampling subjects differs across treatment and control groups, then the ratio of the sample sizes is inversely proportional to the square root of the relative costs. Interestingly, differences in sampling costs have exactly the same effect on relative sample sizes of treatment and control groups as differences in variances.

In those instances where the unit of randomization is different from the unit of observation special considerations must be paid to correlated outcomes. Specifically, the number of observations required is multiplied by $1 + (m - 1)\rho$, where ρ is the intracluster correlation coefficient and m is the size of each cluster. The optimal size of each cluster increases with the ratio of the within to between cluster standard deviation, and decreases with the square root of the ratio of the cost of sampling a subject to the fixed cost of sampling from a new cluster. Since the optimal sample size is independent of the available budget, the experimenter should first determine how many subjects to sample in each cluster and then sample from as many clusters as the budget permits (or until the optimal total sample size is achieved).

A final class of results pertains to designs that include several levels of treatment, or more generally when the treatment variable itself is continuous, but we assume homogeneous treatment effects. The primary goal of the experimental design in this case is to simply maximize the variance of the treatment variable. For example, if the analyst is interested in estimating the effect of treatment and has strong priors that the treatment has a linear effect, then the sample should be equally divided on the endpoints of the feasible treatment range, with *no* intermediate points sampled. Maximizing the variance of the treatment variable under an assumed quadratic, cubic, quartic, etc., relationship produces unambiguous allocation rules as well: in the quadratic case, for instance, the analyst should place half of the sample equally distributed on the endpoints and the other half on the midpoint. More generally, optimal design requires that the number of treatment cells used should be equal to the highest polynomial order plus one.

The remainder of our study proceeds as follows. Section 2 reviews several basic randomization techniques. We summarize how to calculate optimal sample sizes in Section 3. Section 4 elaborates on these considerations, and includes formulae for binomial outcomes and cluster designs. Section 5 discusses sample arrangement when varying treatment lev-

els are possible. Section 6 concludes.

2 Randomization Techniques

One key feature that differentiates empirical approaches within economics is how they formulate the proper counterfactual, or estimate the treatment effect of interest. To provide some formalization, we consider the outcome Y_i of subject i under treatment and control, $T = 0$ and $T = 1$, respectively. We assume that it can be modeled as a function of observable variables X_i, an unobserved person-specific effect α_i, an average treatment effect $\bar{\tau}$,a person-specific treatment effect τ_i, where $E(\tau_i) = 0$, and ε_i, which is assumed i.i.d.

$$Y_{iT} = \alpha_i + X_i\beta + \bar{\tau}T + \tau_i T + \varepsilon_i \tag{1}$$

The average treatment effect can then be defined as:

$$\bar{\tau} = E(Y_{i1} - Y_{i0}) = E(Y_{i1}) - E(Y_{i0})$$

The identification problem is that we can only observe $E(Y_{i1}|T = 1)$ and $E(Y_{i0}|T = 0)$, where $T = 1$ or $T = 0$ for a given i. Because it is impossible to observe unit i in both states (treatment and no treatment), it is necessary to construct a proper counterfactual (i.e., we cannot observe $E(Y_{i1}|T = 0)$ and $E(Y_{i0}|T = 1)$). If the propensity to receive treatment is correlated with any of the unobserved variables, then the estimate of the average treatment effect is biased since

$$\hat{\tau} = E(Y_{i1}|T = 1) - E(Y_{i0}|T = 0) \neq E(Y_{i1}) - E(Y_{i0})$$

The approach used by experimentalists typically achieves identification via randomization. The experimenter randomly assigns units to receive exposure or non-exposure to treatment and then compares the outcomes of units that received treatment to the outcomes of units that did not receive treatment. Randomization ensures that the assignment to treatment is independent of other sources of variation, and that any bias is balanced across treatment and control groups, thus ensuring that the estimate of the average treatment effect is unbiased:

$$\hat{\tau} = E(Y_{i1}|T = 1) - E(Y_{i0}|T = 0) = E(Y_{i1}) - E(Y_{i0}) = \bar{\tau}$$

In this section we discuss how, given that sample sizes in experiments are always of limited size, the experimenter should assign treatment. There is a large statistical literature on this issue, thus we aim to present a succinct overview of the main methods and their advantages and disadvantages. It should be highlighted that our discussion will continue to focus on measuring average treatment effects, which has consumed much of the experimental literature. This is because it is in the spirit of classical experimental design; yet we should note that this leaves important issues on the sidelines, such as heterogeneity of treatment effects (see List, 2006, for a general discussion, and Loomes 2005 and Wilcox 2008 for studies that reveal the repercussions of this choice in measuring expected utility violations). More broadly, we urge caveat lector because in some cases the principles for choosing optimal designs might differ from the principles considered here. Kanninen (2002) provides a beautiful illustration of this fact when the goal is to measure the parameters of a binomial logit model.

2.1 Block and Within Subject Designs

The simplest experimental design is a completely randomized design, where treatments are probabilistically assigned to subjects independent of any of the subject's observed or unobserved characteristics. The advantage of this procedure is that it minimizes the risk that treatment is correlated with individual characteristics. The disadvantage is that the variance of outcomes is potentially very large and the sample sizes of treatment and control groups are randomly generated. Both of these problems reduce the experimenter's ability to draw statistical inference from the experiment.

Instead, if the subject pool is heterogeneous in various dimensions the experimenter may want to reduce the variance of the unobserved component. This can be done subsequent to the experiment by including observable variables X_i in a linear regression and thus constructing an estimate of the average treatment effect with lower variance in finite samples. Alternatively, the conditioning can be built into the design of the experiment. The basic strategy used for incorporating subject heterogeneity into the design of an experiment is to divide the experimental units into blocks. The idea is to treat heterogeneous characteristics of subjects as further treatments. Randomization is within, but not between blocks, thus ensuring that all treatment effects, including the effect of subject characteristics, can be identified. Note that blocking, or equivalently including observable variables in the subsequent regression, will typically decrease the variance of the estimate of the average treatment effect. Specifically, note that

$$var(\hat{\tau}) = \frac{\sigma^2}{N} = \frac{var(\varepsilon)}{N * var(T)} \tag{2}$$

The variance of the estimate of the average treatment effect σ^2/N is increasing in the variance of the unobserved component $var(\varepsilon)$, and decreasing in the number of observations N and the variance of the treatment propensity $var(T)$.[2] Blocking or conditioning on X increases efficiency by reducing the variance of the unobserved component. Another advantage is that blocking allows estimation of an average treatment effect over subsamples of the subject pool. In this case, there is a distinct benefit from blocking prior to the experiment since one can ensure that the standard error of the estimate of treatment effects for each subsample is as small as possible, as discussed below.

A within subject experimental design, in which the same subject experiences more than one experimental treatment, can be thought of as a special case of the block design where the experimenter blocks on a single subject. A main advantage of the within subject design is that it may greatly reduce the variance of the unobserved component, increasing the precision of the estimated average treatment effect. Specifically, assuming that outcomes are generated by equation (1) then, conditional on X, the difference in the variance of the estimate of the treatment in a between subjects and a within subject design is given by:

$$\sigma^2_{BS} - \sigma^2_{WS} = \frac{2}{N} var(\alpha_i)$$

[2] More generally, $Var(\hat{\tau}) = \frac{Var(\varepsilon)}{N * Var(T) * \left(1 - R^2_{XT}\right)}$. But since treatment is assigned at random, X and T are uncorrelated so that R^2_{XT} (the R-squared of a regression of T on X) is equal to zero.

where σ^2_{BS} and σ^2_{WS} are, respectively, the conditional between and within subject variance.[3] In addition, fewer subjects have to be recruited for a within subject design and the degrees of freedom are larger. A disadvantage of the within subject design is that treating a single subject multiple times may result in complicated interactions between treatments and thus yield a different parameter than is estimated in the between experimental design. These context effects include history and learning effects, and sensitization to perceived dependencies across trials (see Greenwald, 1976). Some of these more complicated effects can be controlled for using crossover designs, where the order in which treatments are applied to a subject is randomized. For example, if the outcome is determined by equation

$$Y_{it} = X_{it}\beta + \bar{\tau}T + \tau_i T + \bar{\gamma}T_{(t-1)} + \gamma_i T_{(t-1)} + \varepsilon_i$$

then applying treatment T and control C in the order TC and CT allows for identification of $\bar{\tau}$. More complicated interactions may be identified under a more elaborate TCT and CTC crossover design to achieve identification. However, within subject designs potentially suffer from the problem that treatments may interact in unexpected ways. This issue in and of itself merits an entire study, but we close the discussion urging scholars to take caution when interpreting treatment effects measured using within subject designs.

2.2 Factorial Designs

A completely random or random block design has the disadvantage that sample sizes may vary considerably across blocks. In a factorial design the experimenter chooses a pre-determined number of subjects to each combination of treatments, which can greatly increase the efficiency of the design. Randomization in this case is over the order in which treatments are assigned to experimental units. For example, subjects should not be assigned to treatment and control groups in the order in which they arrive at the laboratory, since early and late arrivals may differ systematically. Instead, each subject should be assigned a random number, based upon which assignment to treatment or control is carried out.

A basic factorial design has the same number of subjects assigned to each combination of treatments. Further, it is likely to be expensive to run all possible combinations of treatments: with n treatments this would require $2n$ trials. However, in the absence of interaction effects between treatments, only $n + 1$ trials are necessary to identify all treatments effects. These $n + 1$ trials must be linearly independent to guarantee that all treatment effects can be identified. The advantage of this fractional factorial design approach is a reduced number of trials. A major disadvantage is that in its simplest form

[3] The within-subject design, however, does not in general have to result in a lower variance of the estimate of the treatment effect. If we allow for individual fixed effects and the treatment effects to be correlated:

$$Y_{iT} = \alpha_i + X_i\beta + \bar{\tau}T + \tau_i T + \alpha\tau_{ij}T + \varepsilon_i$$

then

$$\sigma^2_{BS} - \sigma^2_{WS} = \frac{2}{n}\left[var\left(\alpha_i\right) - var\left(\alpha\tau_{ij}\right)\right]$$

which is no longer unambiguously positive. See Keren (1993) for a derivation of these results and an overview of factors that influence the choice in between- or within-subjects design.

such an approach renders it impossible to check for the existence of interaction effects. Moreover, as we discuss below, the basic factorial design, with equal sample sizes in each treatment cell, is likely to be inefficient.

As Levitt and List (2009) discuss, one potential problem arising from any randomization approach is "randomization bias," a situation wherein the experimental sample is not representative of the population of interest due to the randomization itself. This problem emanates from the field of clinical drug trials, where it has been found that persuading patients to participate in randomized studies is much harder than persuading them to participate in non-randomized studies (Kramer and Shapiro, 1984). In principle, randomization bias also might influence experiments in economics. In particular, laboratory experiments as well as artefactual and framed field experiments might suffer from randomization bias (see Harrison and List, 2004). The one study that we are aware that explores this issue is the work of Harrison et al. (2009). Using an artefactual field experiment to explore risk preferences, they find that (p. 1): "randomization bias is not a major empirical problem for field experiments of the kind we conducted..." Certainly more work is necessary, but our intuition is that randomization bias will not present itself as a major impediment to measurement in the same manner observed in clinical drug trials.

3 Optimal Sample Arrangement: Basics

Given a randomization scheme an important issue to consider is the optimal sample size in each treatment cell. In calculating optimal sample sizes an experimenter must consider three key elements: (1) the significance level, (2) the power of the subsequent hypothesis test, and (3) the minimum detectable effect size. The significance level of a hypothesis test is the probability of falsely rejecting the null hypothesis (also known as the probability of a Type I error). The power of a statistical test is the probability that it will correctly lead to the rejection of the null hypothesis (the probability of a Type II error is 1 - power, and is equal to the probability of falsely not rejecting the null hypothesis).[4] The effect size is the magnitude of the treatment effect that the experimenter wants to detect.

In this section we derive an explicit formula for experiments that have a dichotomous treatment, where the outcome is continuous and we assume that a t-test will be used to determine differences in means between treatment and control group.[5] The formula illustrates the trade-offs inherent in the choices that experimenters face and we make these more tangible by providing empirical examples. In subsequent sections we consider further cases: binomial outcomes, cluster designs, and varying treatment intensities. We also discuss cases where sampling costs for treatment and control are unequal and where

[4] Discussions of power tend not to be intuitively appealing to economists. This is because our usual approach stems from the standard regression model: under a true null what is the probability of observing the coefficient that we observed? Power calculations are altogether different, exploring the question of: if the alternative hypothesis is true, then what is the probability that the estimated coefficient lies outside the 95% confidence interval defined under the null.

[5] The sample size calculations depend on the hypothesis test the experimenter will ex post employ to analyse the data. For power calculations using non-parametric statistical tests see, for example, Rutstrom and Wilcox (2007).

the cost of an additional subject in a new cluster is not the same as that of a subject in a cluster that has already been sampled. In practice, an experimenter can draw upon statistical software to help calculate sample sizes if different hypothesis tests are to be used.[6]

3.1 Dichotomous Treatment and Continuous Outcome

Using the empirical specification above, a single treatment T results in (conditional) outcomes Y_{i0} if $T = 0$ where $Y_{i0}|X_i \sim N(\mu_0, \sigma_0^2)$ and Y_{i1} if $T = 1$ where $Y_{i1}|X_i \sim N(\mu_1, \sigma_1^2)$. In the model given by equation (1) $\sigma_1^2 - \sigma_0^2 = var(\tau|X)$. Only if the variance of the individual specific treatment effects equals zero, i.e. the treatment effect is homogeneous, will the variances across treatment and control groups be equal. Since the experiment has not yet been conducted, the experimenter must form beliefs about the variances of outcomes across the treatment and control groups, which may, for example, come from theory, prior empirical evidence, or a pilot experiment. The experimenter also has to make a decision about the minimum detectable difference between mean control and treatment outcomes, $\mu_1 - \mu_0 = \delta$, that the experiment is meant to be able to detect. In essence, δ is the minimum average treatment effect, $\bar{\tau}$, that the experiment will be able to detect at a given significance level and power. Finally, we assume that the significance of the treatment effect will be determined using a t-test.

Calculating optimal sample sizes requires specifying a null hypothesis and a specific alternative hypothesis. Typically, the null hypothesis is that there is no treatment effect, i.e. that the effect size is zero. The alternative hypothesis is that the effect size takes on a specific value (the minimum detectable effect size). The idea behind the choice of optimal sample sizes in this scenario is that the sample sizes have to be just large enough so that the experimenter (1) does not falsely reject the null hypothesis that the population treatment and control outcomes are equal, i.e. commit a Type I error; and (2) does not falsely accept the null hypothesis when the actual difference is equal to δ, i.e. commit a Type II error. More formally, if the observations for control and treatment groups are independently drawn and $H_0 : \mu_0 = \mu_1$ and $H_1 : \mu_0 \neq \mu_1$, we need the difference in sample means $\bar{Y}_1 - \bar{Y}_0$ (which are of course not yet observed) to satisfy the following conditions:

1. A probability α of committing a Type I error in a two-sided test, i.e. a significance level of α. This is true if

$$\frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = t_{\alpha/2} \Rightarrow \bar{Y}_1 - \bar{Y}_0 = t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{3}$$

   where σ_T^2 and n_T for $T = \{0, 1\}$ are the conditional variance of the outcome and the sample size of the control and treatment groups.

[6] Useful software and documentation includes Spybrook et al (2008a, 2008b), Lenth (2001, 2006-2009), StataCorp (2007). Note that optimal sample sizes calculated by various software may not match precisely those that can be derived from the formulae in this paper.

2. A probability β of committing a Type II error, i.e. a power of $1-\beta$, in a one-sided test. This is true if

$$\frac{\left(\bar{Y}_1 - \bar{Y}_0\right) - \delta}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} = -t_\beta \Rightarrow \bar{Y}_1 - \bar{Y}_0 = \delta - t_\beta \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{4}$$

Using (3) to eliminate $\bar{Y}_1 - \bar{Y}_0$ from (4) we obtain

$$\delta = \left(t_{\alpha/2} + t_\beta\right) \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{5}$$

It can easily be shown that if $\sigma_0^2 = \sigma_1^2 = \sigma^2$, i.e. $var\left(\tau_i\right) = 0$, then the smallest sample sizes that solve this equality satisfy $n_0 = n_1 = n$ and then

$$n_0^* = n_1^* = n^* = 2\left(t_{\alpha/2} + t_\beta\right)^2 \left(\frac{\sigma}{\delta}\right)^2 \tag{6}$$

If the variance of the outcomes are not equal this becomes

$$\begin{aligned} N^* &= \left(\frac{t_{\alpha/2} + t_\beta}{\delta}\right)^2 \left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right) \\ \pi_0^* &= \frac{\sigma_0}{\sigma_0 + \sigma_1}, \quad \pi_1^* = \frac{\sigma_1}{\sigma_0 + \sigma_1} \end{aligned} \tag{7}$$

where $N = n_o + n_1$, $\pi_o + \pi_1 = 1$, $\pi_0 = \frac{n_0}{n_o + n_1}$.

If sample sizes are large enough that the normal distribution is a good approximation for the t-distribution, then the above equations are a closed form solution for the optimal sample sizes. If sample sizes are small, then n must be solved by using successive approximations. Optimal sample sizes increase proportionally with the variance of outcomes, non-linearly with the significance level and the power, and decrease proportionally with the square of the minimum detectable effect. The relative distribution of subjects across treatment and control is proportional to the standard deviation of the respective outcomes. This suggests that if the variance of outcomes under treatment and control are fairly similar there should not be a large loss in efficiency from assigning equal sample sizes to each.

Equation 6 makes it quite clear that any simple rule of thumb–such as place 30 subjects in each experimental treatment cell–has little basis in terms of power unless the researcher believes that he wants to detect an approximately 0.70 standard deviation change in the outcome variable. More generally, equation 6 can be used as a simple heuristic to compute sample sizes necessary to detect various effects. For example, following the standards in the literature and using a significance level of 0.05, and setting power to 0.80, we have $t_{\alpha/2} = 1.96$ and $t_\beta = 0.84$ from standard normal tables. Thus, one would need n = 16 (64) observations in each treatment cell to detect a one (one-half) standard deviation change in the outcome variable.

3.2 An Empirical Example

A quick perusal of the experimental studies published in the social sciences as well as conducted in the business community makes it clear that the status quo is to attempt to include an equal number of subjects in every experimental cell. The summary above provides a strong reason why we should be careful with this aspect of the design since we might fail to maximize power if we do not consider optimal sample arrangements. Consider List (2001) as one illustrative example. List conducted a valuation field experiment at a sportscard show exploring how agents bid in Vickrey second-price auctions for a baseball card. We focus here on the comparison of two treatments among non-sportscard dealers: hypothetical versus actual bidding distributions. The underlying idea, therefore, is that in this case we might have heterogeneous treatment effects in that agents respond differently to hypothetical auctions.

Indeed, previous work suggests that, in general, valuations in hypothetical settings have a greater variance than valuations in tasks that are monetarily binding (see, e.g., Camerer and Hogarth, 1999). Putting aside the issue of heterogeneous costs to obtain sample points, the design fails to adequately adjust sample sizes for the greater expected variance in hypothetical bids. In the paper, the sample sizes for each group are almost equivalent, while the standard deviation of bids in the hypothetical auction is almost twice the standard deviation of bids in the actual auction.[7] At a ratio of standard deviations of 2:1 the suboptimal design (with equal sample sizes in both groups) requires an 11% larger total sample size than the optimal sample design (with the ratio of sample sizes equal to the ratio of standard deviations) to achieve the same power. Specifically, using equation (7), we calculate that given the total sample $N = 175$, the optimal sample sizes for the hypothetical and actual auction are $n_H = 111$ and $n_A = 64$, respectively. Using a uniform design instead of the optimal one decreases the power of the experiment (at the observed effect size) from 69% to 66%. Had the variances been even more different, the efficiency loss due to non-optimal sample arrangements would have been much larger. All else equal, for a ratio of standard deviations of 3, 4, and 5 the required total sample size in the suboptimal (equal sample size) design is 25%, 36%, and 44% larger than in the optimal design. Similarly, we find that (using equation (5)) the minimum detectable effect size is 12%, 17%, and 20% higher in the suboptimal design. However, with a level of power of 69%, the optimal design is still underpowered relative to the conventional standard of 80%.

3.3 Treatments with Unequal Costs

This far we have implicitly assumed that sampling costs for treatment and control groups are equal. Determining optimal sample sizes is somewhat more complicated upon relaxation of this assumption. For example, in many cases treatment might be more expensive to administer because it is costly to provide the good or service in question. In this case, the key idea remains the same–we want to maximize the minimum detectable effect size

[7] The mean bids (standard deviations) are \$49.03 (\$79.96) and \$25.60 (\$46.23) in the hypothetical and actual auctions respectively. The book value of the Cal Ripken Jr. 1982 Topps Traded baseball card was in the range of \$200-\$250.

as given by equation (5), but now we must consider the cost of applying control and treatment, c_0 and c_1. By maximizing the minimum detectable effect, as given by equation (5), subject to $c_0 n_0 + c_1 n_1 = M$ we find that

$$\frac{n_1^*}{n_0^*} = \frac{\pi_1^*}{\pi_0^*} = \sqrt{\frac{c_0}{c_1}}\frac{\sigma_1}{\sigma_0}$$

As before, the optimal sample sizes are proportional to the standard deviations of the respective outcomes and, in addition, they are inversely proportional to the square root of the relative sampling cost. Hence, if sampling costs for the control group are smaller than for the treatment group, as is frequently the case, then the control group should be larger than the treatment group. Yet, as with unequal variances, since the optimal sample sizes are proportional to the square root of the cost of sampling this only becomes important when the difference in costs grows large.

3.4 Parameter Uncertainty

In estimating optimal sample sizes an experimenter needs to decide on a significance level, power and estimable effect size. The choice of significance level is given by convention at 5%, but deciding on the relevant power is more difficult. Experimenters typically want to reject the null hypothesis that the treatment effect is zero, where the probability of such a rejection is given by the power.[8] For example, running an experiment with a power of 80% means that 20% of the time the experimenter will ex ante not be able to reject the null hypothesis of a zero treatment effect despite there being a significant effect in the population.

Traditionally, economists specify all aspects of an experiment's design in advance of actually beginning the experiment (or at least they claim to do so). However, the major difficulty in obtaining reasonable estimates of optimal sample sizes is that information on the variance of outcomes may be poor. The use of historical data and previous similar experiments are likely to be important sources of information. Frequently, though, it is necessary to conduct a pilot experiment to obtain reasonable estimates of the population parameters. This information is then used in deciding how to design and apply treatments, as well as in deciding the number of subjects to be sampled. The advantage of this approach is that the results can be analyzed using the typical parametric or non-parametric significance tests, with the usual p-values. The disadvantage of this approach is that the experimental design cannot be adapted as new information is revealed. As the experiment progresses the experimenter may realize that initial estimates of the optimal

[8] In cases where the experimenter is interested in the non-rejection of the null hypothesis equivalence testing is useful. Failure to reject a null hypothesis does not provide unequivocal evidence that there is no treatment effect, since the failure to reject may actually be the result of low statistical power. In equivalence testing, the researcher decides on a value Δ, where if the effect size is no larger than that value it can be considered negligible. Thus, the null hypothesis becomes that a treatment has a large effect, or $H_0 : |D| > \Delta$, where D is the actual treatment effect. The alternative hypothesis is $H_q : |D| < \Delta$. The equivalence test entails two one-sided α level hypothesis tests. Schuirmann (1987) shows that if a $1 - 2\alpha$ confidence interval lies entirely between $-\Delta$ and Δ, then we can reject the null hypothesis in favor of equivalence at the α level.

sample size may have been too large or too small (indeed, due to the randomness inherent in sampling and due to poor initial estimates of key parameters, such as the variance of the unobserved component σ^2, this is likely to be the case). So, for example, at the end of an experiment the relevant p-value may turn out to be 6%. One approach is to say that the trial was "underpowered" and another trial should be carried out. Alternatively, some argue that it would be more efficient to simply increase the sample size in the present trial with the goal of resolving the issue. The problem is that an experimental approach in this spirit increases the Type I error rate.

One solution is to conduct group-sequential trials combined with interim analysis. This requires the experimenter to conduct the experiment with groups of subjects of predetermined size. After each group the data is analyzed, using cutoff p and t-values that have been adjusted depending on the exact procedure used, and then the decision is made as to whether to continue or not. The two most popular such designs are due to Pocock (1977) and O'Brien-Fleming(1979) (see Lewis,1993, for an accessible introduction to the use of interim analysis).These methods may, however, be inefficient (e.g. in terms of required statistical corrections) and impractical (e.g. it is difficult to verify ex post that an experiment followed an ex ante procedure). An alternative is to use adaptive designs that take a more flexible Bayesian approach to experimental design. A number of such approaches have been explored in the clinical trial literature, including adapting sample sizes and dosage levels during an ongoing clinical trial (see Berry, 2004, for an overview). Hahn et al (2009) develop a "propensity score" method that uses estimates of heterogenous treatment effects from the first stage to set the conditional probability of treatment in the second stage, following the optimal allocation of sample sizes under unequal variances (equation (6)). Further examples in the economics literature include El-Gamal, McKelvey and Palfrey (1993) and El-Gamal and Palfrey (1996). These designs are more difficult to implement, but are especially attractive if the cost of sampling is prohibitively high.

4 Optimal Sample Arrangements: Further Considerations

4.1 Dichotomous Treatment and Binomial Outcomes

The formulae for the continuous case in a between subject design can be adapted for other common experimental designs including: within subject designs, cluster designs, and binary outcomes. As in the continuous case, we assume that we can use the normal approximation to the given distribution. We then substitute into the equations above the appropriate variance estimates for the distribution of interest. Note that in the cases of binary and count data, the variance depends on the mean. Thus, in equation (3), under which the null hypothesis is true, the treatment and control groups will have equal means and therefore equal variances and equal optimal sample sizes. In equation (4), under which the alternative hypothesis is true, the treatment and control groups will have different means and therefore different variances. For binary data, using the normal approximation to the binomial distribution, the variance is equal to $p(1-p)$ where p is

the mean of the outcome variable. For a null hypothesis $H_o : p_0 = p_1$ (proportions are equal under treatment and control) the optimal sample sizes are equal

$$n_0^* = n_1^* = n^* = \left(t_{\alpha/2}\sqrt{2\bar{p}\left(1-\bar{p}\right)} + t_\beta\sqrt{p_0\left(1-p_0\right)+p_1\left(1-p_1\right)}\right)^2 \delta^{-2} \tag{8}$$

where $\bar{p} = \left(p_0 + p_1\right)/2$.[9]

Because the variance $p(1-p)$ will be maximized for $p = 0.5$, optimal sample sizes will increase as $\bar{p}$ approaches 0.5 (i.e., sample sizes decrease in $|\bar{p} - 0.5|$). Similarly, if the null hypothesis is of the form $p_1 = kp_0$, where $k > 0$, then the sample size arrangement is dictated by k in the same manner as in the continuous case using equation (7). The closer p_1 is to 0.5 relative to p_0, the larger the proportion of the total sample size that should be allocated to p_1 (and vice versa).

4.2 Cluster Designs

Thus far we have assumed that the unobserved components are independently distributed among subjects. However, in particular with the recent growth in field experiments, the possibility of correlation in the unobserved component among subjects within a cluster needs to be considered. Some recent field experiments that emulate social experiments commonly feature cluster randomization, in which clusters of individuals rather than independent individuals are randomly allocated to intervention groups. A key property of cluster randomization trials is that the outcome of interest may occur at the individual level whereas the randomization occurs at the cluster or group level. Thus, the unit of randomization is different from the unit of statistical analysis. For example, an intervention aimed at improving individual health might be randomly assigned to villages. In this case, the lack of independence among individuals in the same village will affect both the optimal sample sizes and the analysis of the experimental results. As we illustrate in the example below, the adjustment to sample sizes due to clustering can be substantial.

Consider the case where each subject is also a member of a group j and outcomes for $T = \{0, 1\}$ are given by

$$Y_{ijT} = \alpha + \bar{\tau}T_j + \nu_j + \varepsilon_{ij}$$

with ε_{ij} the individual specific i.i.d. error term and ν_j a group specific i.i.d. error term (we ignore X_i, α_i and τ_i for simplicity). Suppose that sampling is by cluster, where each cluster is of size m for both treatment and control groups. Under the assumption of equal variances across treatment and control groups (and thus equal sample sizes), optimal sample sizes in cluster designs can be calculated via the following equation:

$$n_0^* = n_1^* = n^* = 2\left(t_{\alpha/2} + t_\beta\right)^2\left(\frac{\sigma}{\delta}\right)^2\left(1 + (m-1)\rho\right) \tag{9}$$

with $2(k-1)$ degrees of freedom (assuming no other covariates), where $k = \frac{n}{m}$ is the number of clusters, σ^2 is the common variance of the treatment and control groups without

[9] Note that the above equation (8) follows the null hypothesis of equal means (and thus variances) for sample sizes and the significance test; and follows the alternative hypothesis of different means (and thus variances) for the power test. See Fleiss et al (2003) for further discussion of binary data.

clustering, and $\rho = \frac{var(\nu_j)}{var(\nu_j)+var(\varepsilon_{ij})}$ is the coefficient of intracluster correlation. This is simply our previous expression, as given by equation (6) augmented by the "variance inflation factor", $1 + (m-1)\rho$ (see Donner and Klar, 2000, for further discussion of this result). Equation (9) shows that the necessary total sample size in a cluster design increases (near) proportionally with both the size of each cluster and the intracluster correlation. Also notice that the degrees of freedom in a cluster design are far smaller, further increasing the necessary sample size. Hence, in the presence of intracluster correlation, $\rho \neq 0$, it is important, if possible, to randomize over as small clusters as possible so as to maximize the efficiency of the experiment.

Figure 1 plots the power of two cluster designs for a given intracluster correlation ρ and a standardized effect size δ/σ equal to 0.2. Two scenarios are considered: clusters of fixed size ($m = 20$) with the number of clusters k allowed to vary; and, a fixed number of clusters ($k = 20$) with cluster size m allowed to vary. As shown in the figure, the power of the fixed number of clusters design quickly flattens out so that adding an additional person to each cluster yields little gain in power. Whereas adding the same number of subjects to the study in the form of a new cluster (i.e., in the clusters of fixed size design) yields relatively large power gains. For example, a sample size $n^* = 1,000$ can be allocated to 20 clusters with 50 subjects each (under the fixed number of clusters design) or to 50 clusters with 20 subjects each (under the fixed size of cluster design) yielding power levels of 75% and 45% respectively ($\rho = 0.1$). As the figure also illustrates, the loss of power due to an in increase in ρ is substantial.

The decision on the optimal number of clusters k and number of subjects in each cluster m in a cluster design will depend on the cost of sampling within a cluster and the fixed cost of starting to sample from a new cluster. Denoting c_m as the cost of each subject and c_k as a fixed cost per cluster, then the total cost of collecting the data is $2\left(c_m m + c_k\right) k = M$. Maximizing the minimum detectable effect size, found by rearranging equation (9), subject to this budget constraint yields an expression for the optimal size of each cluster

$$m^* = \sqrt{\frac{(1-\rho)}{\rho}}\sqrt{\frac{c_k}{c_m}} \tag{10}$$

where $\frac{(1-\rho)}{\rho} = \frac{var(\varepsilon_{ij})}{var(\nu_j)}$. The optimal cluster size is proportional to the square root of the ratio of the fixed cost per cluster and the cost per subject, and to the ratio of the standard deviation of the within and between cluster variation. Perhaps surprisingly, the optimal cluster size is independent of the total budget available for the experiment and thus on a limited budget the experimenter should first work out how many subjects to sample from each cluster and then sample as many cluster as is affordable. The optimal number of clusters k^* is found by substituting the expression for the optimal cluster size m^* from equation (10) back into equation(9), recalling that $n = mk$.[10]

[10]See, for example, Bloom (2005), Donner and Klar (2000), Martinez et al (2007), Raudenbush (1997) and Spybrook et al (2008b) for a further discussion of optimal cluster design.The software available in Spybrook (2008a) and documented in Spybrook et. al. (2008b) is a comprehensive tool for designing cluster level experiments.

5 Optimal Sample Arrangement: Varying Treatment Levels

5.1 Varying Treatment Levels and Continuous Outcomes

This section explores optimal design when the treatment variable is permitted to take on varying levels. For example, the analyst wants to explore different price levels or different dosage levels under the assumption of no heteroscedasticity (i.e., homogeneous treatment effects). The reader who is interested in cases of varying treatment levels and variance heterogeneity should see Kish (1965) and Wilcox (1996).

To begin, let us return to the empirical specification above (equation (2)), but now consider the simpler case where $\tau_i = 0$ for all i; thus treatment and control outcomes have the same variance. Now outcome Y_i is a function of observable variables X_i, a linear function of the treatment variable T_i and ε_i, which is assumed i.i.d.

$$Y_i = X_i\beta + \bar{\tau}T_i + \varepsilon_i$$

The goal in this case is to derive the most precise estimate of $\bar{\tau}$ by using exogenous variation in T. To add further structure to the problem, we assume that the outcome variable is measurable in continuous units (binary data outcomes do not change the nature of the arguments) and the experimenter can set the treatment variable over the range $\left[0, \overline{T}\right]$.

Each time we present this type of exercise to our students, querying them about the optimal sample arrangement, the modal response is one of uniformity: either "split the sample into integers and equally distribute the sample," or "split the sample into equivalently sized cells" naturally become the crowd favorites. Before considering the correct response, similar to the case with dichotomous treatment, it is useful to reflect on the mechanics of the regression model in the relationship given above. To maximize precision, one must first consider techniques to minimize the variance of the estimated treatment effect. Recall that $var\left(\hat{\tau}\right) = \frac{var(e)}{n*var(T)}$. This simple relationship provides three ways to increase precision: (1) decrease the variance of the unobserved component $var\left(\varepsilon\right)$, (2) increase the sample size n, or (3) increase the variance of the treatment effect $var\left(T\right)$. We are struck by the fact that in most literatures, including our own, discussions surrounding changes in sample size, perhaps the costliest approach, dominate the landscape when considering techniques to increase precision. Yet, there is an exact trade-off inherent in experimental design that is clear from the regression model. For example, tripling the variation in T has an identical effect on precision as tripling the sample size.

If the experimenter has strong priors that the effect of treatment is linear, then it is straightforward to see that the variance of T is maximized by placing half of the sample in treatment cell $T = 0$ and half of the sample in treatment cell $T = \bar{T}$. Clearly, this maximizes the variance of the treatment effect and hence minimizes the standard error of the estimate of the treatment effect (for derivations of this and the following results we direct the reader to Atkinson and Donev, 1992, and Mead, 1988). Hence, the optimal sample design if a linear treatment effect is to be identified is to place half of the sample at each of the extremes of the range of potential treatment intensities. The overall sample size can then be calculated using equation (6) where σ^2/n is given by equation (2).

If the analyst believes that the intensity of treatment T has a non-linear effect on the outcome variable, then clearly sampling from two values of T is inappropriate since non-linear effects cannot be identified. In general, identification requires that the number of treatment cells used should be equal to the highest polynomial order plus one. For example, if a quadratic relationship is presumed, then three treatment cells should be chosen in the feasible range. Further, in this case those treatment cells selected should be at the extremes and at the midpoint of the range, $T = \left\{0, \bar{T}/2, \bar{T}\right\}$, where the optimal proportions in each of these treatments cells is $\left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$. As McClelland (1997) notes, intuitively the test for a quadratic effect compares the mean of the outcomes at the extremes to the mean of the outcome at the midpoint; and as before, the variance is maximized when equal proportions are allocated to the these two means: the midpoint and the extremes (and the observations at the extremes are also equally divided). If both a linear and a quadratic effect are included, then the problem becomes considerably more complicated, with the solution being a weighted average of the linear and quadratic optimal allocations (see Atkinson and Donev, 1992).[11]

A related problem is one where the treatment levels are ordinal (categorical) rather than continuous. In this situation it is key to decide which contrasts are of primary interest. For example, take a situation where there are three treatment scenarios $\{A, B, C\}$. Imagine the researcher is primarily interested in comparing outcomes under a baseline scenario A with outcomes under two alternative scenarios $\{B, C\}$, but the outcomes under scenarios B and C will not be compared with each other. In that case the optimal allocation weights more heavily toward A, $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right\}$, since, intuitively, scenario A is used in two contrasts. If instead the mean difference in outcomes under B and C is of primary interest then the optimal allocation is $\left\{0, \frac{1}{2}, \frac{1}{2}\right\}$. The interested reader should see McClelland (1997) for a more detailed discussion.

5.2 An Empirical Example

Similar to the example provided above, in this section we discuss an empirical example that illustrates our points by assuming a significance level of 5% and a power level of 80%. We also assume a fixed experimental budget. Our chosen example is in the area of the economics of charity. Recently a set of lab and field experiments have lent insights into the "demand side" of charitable fundraising. In this spirit, Karlan and List (2007) designed a natural field experiment to measure key parameters of the theory. In their study, they solicited contributions from more than 50,000 supporters of a liberal organization. They randomized households into several different groups to explore whether upfront monies used as matching funds promotes giving. Among other things they tested whether larger match ratios induced more giving. In particular, they use three treatment cells corresponding to match ratios of 3:1 (i.e., \$3 match for every \$1 donated), 2:1 and 1:1. Above we argued that if one were merely interested in estimating a linear price effect

[11] As discussed in McClelland (1997), the optimal allocation for quadratic effects $\{\frac{1}{4},\frac{1}{2},\frac{1}{4}\}$ yields a relative efficiency (in terms of variance) for linear effects of 0.5 compared to the optimal allocation $\{\frac{1}{2},0,\frac{1}{2}\}$. Equal allocation across the three groups $\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$ yields a relative linear efficiency of 0.67 and a relative quadratic efficiency of 0.89 A compromise design $\{\frac{3}{8},\frac{1}{4},\frac{3}{8}\}$ yields relative efficiencies to 0.75 for both linear and quadratic effects relative to optimal allocations.

over this range, and suppressed cost considerations, then the 1:1 and 3:1 cells should have been the only ones sampled. Given a fixed number of subjects, we use equation (5) to calculate that the minimum detectable effect of the three treatment cells design (with an equal distribution of subjects across treatment cells) is about 22% higher than that of the two treatment cell design.[12]

Alternatively, for a given power, significance level and minimum detectable effect the three treatment cell design requires, using equation (6), 50% more observations than the two treatment cell design. Suppose, instead of a linear effect, the authors were interested in estimating a quadratic effect and had allocated the sample accordingly (i.e., half the sample in the 2:1 cell and one-quarter of the sample each in the 1:1 and 3:1 cells). If in fact the treatment effect turned out to be linear rather than quadratic, this design would result in a minimum detectable effect that is about 41% higher than that of the optimal two treatment cell design for linear effects. Of course, in practice the authors had to consider the cost of sampling, but this example illustrates the gains under the assumption of cost symmetry.

6 Concluding Remarks

In experimental economics discussion of optimal sample size arrangement is rare. In this way, finding a study that makes use of the rules of thumb discussed herein is akin to a ballroom dancer searching for a partner in a hip-hop dance club. Of course, there are good reasons that we are hip-hoppers. First, the effect size and variance are both unknown and difficult to guess without robust data, which could be costly to collect. Second, the analyst might be involved in multiple hypothesis testing, and employing a multiple treatment design allows the analyst to avoid the case of one insignificant result by using a series of weak tests, making it highly likely that a statistically significant result will emerge. Third, the status quo is powerful: one can readily guess the nature of the referee reports for a paper in which the author chooses to sample only the endpoints of the feasible treatment region. Even in those cases where the referee agrees that linearity is appropriate, we suspect that the referee will be more comfortable with some mid-range sampling. We hope that this study begins to break that mold, and induces experimenters to design more efficient experiments.

In this respect, under a certain set of assumptions this study pinpoints several rules of thumb that experimenters might find useful:

A. With a continuous outcome measure one should only allocate subjects equally across treatment and control if the sample variances of the outcome means are expected to be equal in the treatment and control groups.

B. In those cases where the sample variances are not equal, the ratio of the sample sizes should be set equal to the ratio of the standard deviations.

C. If the cost of sampling subjects varies across experimental cells, then the ratio of the sample sizes is inversely proportional to the square root of the relative costs.

D. When the unit of randomization is different from the unit of analysis, the

[12] We use the estimated standard error of 0.049 from the empirical example (Karlan and List (2007) ,Table 2A, Panel A, col (4)).

intracluster correlation coefficient should be considered.

E. When the treatment variable itself is continuous, the optimal design requires that the number of treatment cells used should be equal to the highest polynomial order plus one. For instance, if the analyst is interested in estimating the effect of treatment and has strong priors that the treatment has a linear effect, then the sample should be equally divided on the endpoints of the feasible treatment range, with *no* intermediate points sampled.

Clearly, this study represents only the tip of the iceberg when it comes to discussing optimal experimental design. For instance, we can imagine that an entire set of papers could be produced to describe how to design experiments based on power testing and confidence intervals. More generally, we hope that methodological discussion eventually sheds its perceived inferiority in experimental economics and begins to, at least, ride shotgun in our drive to a deeper understanding of economic science. Several prominent discussions remain to be heard: generalizability of results across domains (but, see Levitt and List, 2007, and subsequent studies), use of the strategy method, one-shot versus repeated observations, elicitation of beliefs, within versus between subject experimental designs, using experiment to estimate heterogeneous treatment effects; and in the design area more specifically, optimal design with multiple priors and Bayesian and frequentist sample size determination are but just a few areas not yet properly vetted in the experimental economics community.

References

[1] Atkinson, A. C. and Donev, A. N. (1992). Optimum experimental designs. Oxford, England: Clarendon Press.

[2] Berry, D. (2004) "Bayesian Statistics and the Efficiency and Ethics of Clinical Trials." *Statistical Science*, 19:175-187.

[3] Bloom, Howard S. (2005), "Randomizing Groups to Evaluate Place-Based Programs." In Howard S. Bloom, ed., *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation Publications.

[4] Blundell, R. and Costas Dias, M. (2002) "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal*, 1: 91-115.

[5] Camerer, Colin F. and Robin M. Hogarth (1999). "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 7-42.

[6] Cochran, W.G. and Cox, G.M. (1950). Experimental Designs. New York, NY: J. Wiley & Sons, Inc.

[7] Cohen, Jacob (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: L. Erlbaum Associates.

[8] Donner, Allan and Neil Klar (2000). Design and Analysis of Cluster Randomization in Health Research. Holder Arnold, London.

[9] El-Gamal, M., McKelvey, R. and Palfrey, T. (1993). "A Bayesian Sequential Experimental Study of Learning in Games," *Journal of the American Statistical Association*, 88: 428-435.

[10] El-Gamal M. and Palfrey T. (1996). "Economical Experiments: Bayesian Efficient Experimental Designs,"*International Journal of Game Theory*, 25: 495-517.

[11] Fisher, R.A. (1935). The Design of Experiments. Edinburgh, Scotland: Oliver and Boyd.

[12] Fleiss, Joseph L., Bruce Levin and Myunghee Cho Paik (2003). Statistical Methods for Rates and Proportions. Wiley-Interscience: Hoboken, N.J.

[13] Goodman S.N. and Berlin J.A. (1994). "The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results," *Annals of Internal Medicine*; 121:200-6.

[14] Greenwald, A.G. (1976). Within-Subjects Design: To Use or not to Use. Psychological Bulletin, 83, 314-320.

[15] Hahn, J., K. Hirano and D. Karlan (2009). "Adaptive Experimental Design Using the Propensity Score," forthcoming in *Journal of Business and Economic Statistics*.

[16] Harrison, Glenn W. & Morten I. Lau & Elisabet Rutström, E.(2009). "Risk attitudes, randomization to treatment, and self-selection into experiments," *Journal of Economic Behavior & Organization*, Elsevier, vol. 70(3), pages 498-507.

[17] Harrison, Glenn and John A. List (2004). "Field Experiments," *Journal of Economic Literature*, XLII (December): 1013-1059.

[18] Karlan, D. and List, J.A. (2007). "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review*, 97:1774-1793.

[19] Keren, G. (1993). Between or Within-subjects design: A Methodological Dilemma. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological Issues (pp. 257-272). Hillsdale, NJ: Eribaum.

[20] Kish, L. (1965). Survey sampling. New York: Wiley.

[21] Lenth, R. V. (2006-2009). Java Applets for Power and Sample Size [Computer Software]. Retrieved May, 2009 from http://www.stat.uiowa.edu/ rlenth/Power.

[22] Levitt, Steven D. and John A. List (2007). "What do Laboratory Experiments Measuring Social Preferences tell us about the Real World," *Journal of Economic Perspectives*, 21 (2): 153-174.

[23] Lewis, Rojer J. (1993). "An Introduction to the Use of Interim Data Analyses in Clinical Trials," *Annals of Emergency Medicine*, 22(1):1463-1469.

[24] List, John A. (2001). "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *American Economic Review*, 91:5, 1498–1507.

[25] List, John A. (2006). "Field Experiments: A Bridge Between Lab and Naturally Occurring Data," *Advances in Economic Analysis and Policy*, 6(2), Article 8.

[26] Martinez, A., S.W. Raudenbush and J. Spybrook (2007). "Strategies for Improving Precision in Group-Randomized Experiments," *Educational Evaluation and Policy Analysis*, 29(1): 5-29.

[27] Mead, R. (1988). The design of experiments: Statistical principles for practical application. Cambridge, England: Cambridge University Press.

[28] McClelland, Gary H. (1997). "Optimal Design in Psychological Research," *Psychological Methods*, 2(1), 3-19.

[29] O'Brien , P. C. and T. R. Fleming (1979), "A Multiple Testing Procedure for Clinical Trials,".*Biometrics*, 35, 549-556.

[30] Pocock, Stuart J. (1977). "Group Sequential Methods in the Design and Analysis of Clinical Trials." *Biometrika*, 64(2): 191-199.

[31] Raudenbush, S.W. (1997). "Statistical Analysis and Optimal Design for Cluster Randomized Trials," *Psychological Methods*, 2(2): 173-185.

[32] Rutstrom, E. and Wilcox, N. (2007). "Stated Beliefs Versus Inferred Beliefs: A Methodological Inquiry and Experimental Test", *Games and Economic Behavior*, forthcoming 2009.

[33] Schuirmann, D. J. (1987), "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

[34] Spybrook, J., S.W. Raudenbush, X. Liu, R. Congden and. A. Martinez (2008a), Optimal Design for Longitudinal and Multilevel Research v1.76 [Computer Software]. Retrieved last May, 2009. Available at: http://sitemaker.umich.edu/group-based/optimal_design_software

[35] Spybrook, J., S.W. Raudenbush, X. Liu, R. Congden and. A. Martinez (2008b), "Optimal Design for Longitudinal and Multilevel Research: Documentation for the 'Optimal Design' Software," Working Paper. Available at: http://sitemaker.umich.edu/group-based/files/od-manual-20080312-v176.pdf

[36] StataCorp (2007). Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

[37] Wilcox, R. R. (1996). Statistics for the Social Sciences. San Diego, CA: Academic Press.

Figure 1. Power in Cluster Designs