



Laboratory experiments can pre-design to address power and selection issues

Weili Ding¹

Received: 16 April 2019 / Revised: 24 May 2020 / Accepted: 1 July 2020 / Published online: 13 July 2020
© Economic Science Association 2020

Abstract

In this paper, motivated by aspects of preregistration plans we discuss issues that we believe have important implications for how experiments are designed. To make possible valid inferences about the effects of a treatment in question, we first illustrate how economic theories can help allocate subjects across treatments in a manner that boosts statistical power. Using data from two laboratory experiments where subject behavior deviated sharply from theory, we show that the ex-post subject allocation to maximize statistical power is closer to these ex-ante calculations relative to traditional designs that balances the number of subjects across treatments. Finally, we call for increased attention to (i) the appropriate levels of the type I and type II errors for power calculations, and (ii) how experimenters consider balance in part by properly handling over-subscription to sessions.

1 Introduction

To help improve research transparency a number of initiatives including pre-analysis plans, hypothesis registries, and replications have recently sprung into operation across multiple disciplines including political science, neuroscience and economics. These initiatives have been announced by specific academic journals, professional associations and funding agencies to deal with concerns ranging from specification search and failure to replicate.¹ Coffman and Niederle (2015) recently summarize

¹ A partial list of journals which recently introduced a process that would pre-accept papers based on the submission of a detailed proposal for a prospective empirical study prior to results being available include *Psychological Science*, *Journal of Experimental Political Science* and the *Journal of Development Economics*. Several funders of academic research are signatories to the Transparency and Open-

I would like to thank one anonymous reviewer, the guest editor John Ham and Steven Lehrer for many helpful comments and suggestions that have substantially improved the manuscript. Steven Lehrer also generously provided the experimental data analyzed in the study. I wish to thank SSHRC for research support. I am responsible for all errors.

✉ Weili Ding
dingw@queensu.ca

¹ Queen's University, Kingston, Canada

the pros and cons of these initiatives and claims that the benefits of preanalysis plans are likely limited for experimental economists, since there is a strong opportunity for replication by recruiting new subjects.² Our paper will make a case that the reverse is true even for lab experimenters by showing the statistical benefits from aspects involved in pre-designing the study. After all, a central component of a pre-specified analysis plan is that it can allow researchers to take advantage of all the statistical power of well-designed statistical tests and to reduce concerns related to robustness to specifications.

To make this case we consider laboratory experiments that adopt a between-subject design to inform on economic theories.³ The general design of such an experiment is to create a controlled environment in the lab where a single specific theoretical parameter varies between sessions, thereby providing a direct test of some comparative static prediction(s) of a theory.⁴ Similar to randomized controlled experiments in medicine and other fields within economics, using a between-session design is intuitive and easy to explain. Researchers generally focus their interpretation of findings from their experiments through the lens of the outcome of hypothesis tests that compare subject behavior between treatments.

In this paper, we show that thoughtful design of a laboratory experiment may present a host of potential benefits to lab experimenters by improving the overall experimental design prior to its implementation. These benefits are not related to concerns such as “p-hacking” or specification searching, which could be handled by carrying out additional sessions, nor motivated by multiple testing issues.⁵

This paper complements List et al. (2011) and Czibor et al. (2019), who each provide concrete guidance on how to improve lab studies. We extend their guidance in three distinct ways. First, we show how ex-ante statistical power calculations can utilize economic theories to inform on optimal research designs. Second, we discuss

Footnote 1 (continued)

ness Promotion (TOP) Guidelines (<https://cos.io/our-services/top-guidelines/>). Last, among professional organizations, the American Economics Association has created a registry for all randomized controlled trials.

² Related, (Camerer et al. 2016) and (Maniadis et al. 2017) also cast doubt on the value of research transparency initiatives for experimental economists. They conclude from their literature surveys that selective reporting of statistically significant results occurs less frequently in experimental economics relative to other applied fields in economics.

³ Roth (1986) summarizes other possible objectives for laboratory experiments including investigating anomalies and pilot testing of policies. In many other applied fields of economics, policy concerns are the primary motivation to undertake an experiment. Further, lab studies could alternatively employ a within subject design, where the treatment in a session could vary over time allowing additional control for an individual subject specific fixed effect. (Charness et al. 2012) argues that there is a threat to internal validity from sequentially exposing subjects to different treatments, since it may cause order effects.

⁴ For example, to test the predictions of the Abreu and Gul (2000) bargaining theory with obstinate behavioral types, (Embrey et al. 2015) vary the number of behavioral types between sessions.

⁵ To avoid temptations for data mining, in a pre-analysis plan one can list the multiple hypotheses that will be tested and the statistical tests for each hypothesis would be adjusted for multiple inference. See (List et al. 2019) for a detailed discussion targeting experimental economics. Among others, Ding and Lehrer (2011) point out that making statistical corrections for these issues is important in practice whenever there is an opportunity to select the most favorable results from an analysis.

the challenge presented by oversubscription of subjects into a session, an issue that has received little attention in the experimental literature. This final contribution extends Slonim et al. (2013) by making an additional recommendation on how to reduce potential sources of selection bias that could arise during the implementation of the experiment itself.

The rest of the paper is organized as follows. In the next section, we review how statistical power is calculated and discuss how lab experimenters interested in testing economic theories can use knowledge of these models to help their experimental design. We illustrate this approach with data from two experimental studies in which subject behavior deviated greatly from the quantitative predictions of the underlying theory. We find that the ex-post subject allocation to maximize statistical power is closer to these proposed by ex-ante calculations based on the underlying economic theories as compared to traditional designs that simply balance the number of subjects across treatments.⁶ Section 3 examines how pre-analysis plans can further help determine how many sessions to undertake, in which sessions specific treatments should be carried out, and why having a plan to handle over-subscription of subjects to an individual session can be valuable. The final section concludes and suggests a direction for future research.

2 Power analysis for laboratory experiments

2.1 Setting the stage

For expository purposes, we consider a laboratory experiment that compares subject's decision-making across two conditions. We define treatment status for subject i in session s , by an indicator variable D_{is} , where $D_{is} = 1$ denotes random assignment to treatment 1 and $D_{is} = 0$ reflects the status-quo condition (or control). To estimate the effect of the treatment on a continuous outcome y_{ist} measured for subject i at experimental period t in session s , researchers estimate either

$$y_{ist} = \gamma_0 + \gamma_1 D_{is} + e_{1ist} \quad (1)$$

or

$$y_{ist} = \beta_0 + \beta_1 D_{is} + \beta_2 X_{ist} + e_{2ist}, \quad (2)$$

where covariates X_{ist} are included in Eq. (2). These covariates may include demographic information as well as those variables suggested by theory. Both e_{1ist} and e_{2ist} are random error terms. Since D_{is} is randomly assigned, the inclusion of X_{ist}

⁶ Nikiforakis and Slonim (2015) require the use of post-hoc power analysis to identify if the reason a replication failed is due to the study having insufficient power. Heonig and Heisey (2001) point out that in many other scenarios post-hoc power analysis does not aid in the interpretation of p -values from experimental results. This arises since ex post computing the observed power is only possible after observing the p -value and hence cannot change the interpretation of the p -value. Further, they reinforce the need to make power calculations ex-ante to improve the planning of an experiment.

should not affect the expected value of an OLS estimator of the coefficient on treatment between Eqs. (1) and (2), that is $E[\hat{\gamma}_1] = E[\hat{\beta}_1]$, but can influence the size of the standard error on $\hat{\beta}_1$ by reducing the variance of the residual.⁷

Often guided by an underlying theoretical model, experimenters want to test the effect of the treatment D_{is} on a subject outcome of interest. This involves testing some hypothesis about our parameter of interest γ_1 in Eq. (1), or β_1 in Eq. (2). The probability of a Type I error is the probability of rejecting the null hypothesis when it is correct and the probability of a Type II error is the probability of failing to reject the null hypothesis when it is false.

Power analysis is used to optimize hypothesis tests and experimental designs. Since asymptotic normal approximations are valid for tests on $\hat{\gamma}_1$ in a wide variety of applications, we will focus our discussion on the Wald test. The Wald test uses the test statistic $\frac{\hat{\theta} - \theta_o}{SE}$, where $\hat{\theta}$ and SE are the sample estimated statistic and standard error while θ_o is the parameter value under the Null hypothesis. For example, in the commonly used test of statistical significance $\theta_o = 0$ and statistical software would report a t -statistic for Eq. (1) of $\frac{\hat{\gamma}_1 - 0}{\sigma_{\hat{\gamma}_1}}$.

The power of a statistical test is the probability that the test correctly rejects the null hypothesis and is equal to $1 - \text{Pr}[\text{type II error}]$. Holding all other factors constant, the most powerful test is always preferred. The power of the test depends on (i) sample size through the estimated standard error, (ii) probability of a Type I error, and (iii) the effect size $\hat{\gamma}_1$. To show how these variables are related under simple random sampling we present a series of visualizations in Fig. 1.⁸ The first panel illustrates the distribution of our estimator $\hat{\gamma}_1$ under the Null hypothesis ($\gamma_1 = 0$) for a two sided test which defines the rejection region as the shaded area that captures the probability of making a Type 1 error. We refer to this as the Null distribution and the standard deviation of the distribution is given by the estimated standard error $\sigma_{\hat{\gamma}_1}$. The conventional experimental design first fixes the probability of a Type I error at 5%. The effect size presented in the second panel is assumed to be 2.5 standard deviations away from the Null value. The third panel presents the alternative distribution, which is distributed Normal with the mean taking the value of the effect size and the standard deviation identical to that of the Null distribution $\sigma_{\hat{\gamma}_1}$. Prior to

⁷ An additional rationale for estimating Eq. (2) is that it can guard against chance imbalances in important baseline covariates. That said, with linear models, Fisher (1925) was the first to show that adjustment for baseline covariates can lead to an increase in statistical power. Using simulations, Hernandez et al. (1994) show that when the covariates are highly prognostic, the increase in power is substantial. Yet, this property of reducing residual variance does not always hold with non-linear models such as logistic regression (see e.g. Robinson and Jewell 1991) or proportional hazards regression (see e.g. Ford et al. 1995). We briefly discuss power calculations for non-linear estimators in Sect. 2.4.

⁸ Power calculations are specific to the research design and one can simply scale the solution under simple random sampling (SRS) up or down by the design effect (DE). DE is a relative measure of how the variance of a target statistic differs under the proposed design relative to SRS. Later in the text we consider session effects and point out that the DE for stratified sampling is generally < 1 , meaning that smaller samples are needed to achieve the same power (see e.g. Cochran (1977) for details). We suggest in Sect. 3 that stratified sampling can be achieved for lab experiments if they only randomly sent emails for one type of treatment to one subset of their pool, and the remaining sample received an invitation to the control group.

conducting a study, the researcher is unaware of the effect size and can be faced with a difficult choice on choosing a value. However, for many laboratory experiments one could use a prediction from a comparative static exercise of the underlying theory to make a meaningful and plausible choice for the effect size.

The fourth panel shows how to calculate statistical power, the chance a researcher will be able to detect an effect of the size introduced in the second panel, if there really is one. We place a vertical line at the “significance” cutoff of 1.96 from the Null distribution. The unshaded area under the alternative distribution to the left of this line provide us with the probability that we would fail to reject the null if we had an effect of 2.5. The shaded area under the alternative distribution to the right of the line is our statistical power; thus the greater the shaded area, the greater the statistical power. Lastly, sample size plays a role since it affects the size of the estimated standard error and hence variation in both the Null and alternative distributions. Larger samples often increase statistical power.

To undertake power analysis for a research project involving testing a hypothesis about the parameter, γ_1 , first requires determining what a “Minimal Detectable Effect” is. The minimal detectable effect (γ_1^{MDE}) is the smallest possible value for $\hat{\gamma}_1$ at which the Null hypothesis is rejected for a given significance level. For example, undertaking a two-sided test at a significance level of $\alpha = 0.05$, if a researcher wishes to achieve power at the level of $\kappa = 0.8$, where the effect estimator has a limiting normal distribution,⁹ then $\gamma_1^{\text{MDE}} = (t_{\alpha/2} + t_{(1-\kappa)})\sigma_{\hat{\gamma}_1}$. For a large sample size,¹⁰ plugging in $t_{(1-\kappa)}$ and $t_{\alpha/2}$ as obtained from a standard t -distribution yields $\gamma_1^{\text{MDE}} = (1.96 + 0.84)\sigma_{\hat{\gamma}_1} = 2.8\sigma_{\hat{\gamma}_1}$.¹¹ Once a minimal detectable effect is chosen, researchers can rearrange the expressions above to solve for an expression for $\sigma_{\hat{\gamma}_1}$ and achieve such an effect at a specified type 1 error and power level. The expression of $\sigma_{\hat{\gamma}_1}$ depends on two design parameters that the researcher can adjust: sample size and optimal treatment allocation as we demonstrate shortly. Finally, since adding explanatory variables to an estimating equation tends to decrease the residual standard deviation, the required sample size for any specified level of precision or power is thus reduced if estimating Eq. (2) in place of Eq. (1); only if their coefficients are truly non-zero.

2.2 The appropriate choice of α and κ may depend on the audience for the study

In the above example, the values for $\alpha = 0.05$ and $\kappa = 0.8$ were chosen since they reflect the standards commonly used in other branches of social science research. Whether these values should also be the benchmark for studies in experimental economic settings may depend on the goal of the laboratory study. A possible rationale

⁹ A limiting normal distribution is a reasonable approximation so long as sample size is large enough.

¹⁰ If sample sizes are large enough, the normal distribution is a good approximation for the t -distribution.

¹¹ In other words, the minimal detectable effect for a study with these parameters is 2.8 times the standard error of the effect estimator. Examining panel 4 of Fig. 1, we observe that $\hat{\gamma}_1 = 2.5$ so this estimate has lower statistical power than $\kappa = 0.8$.

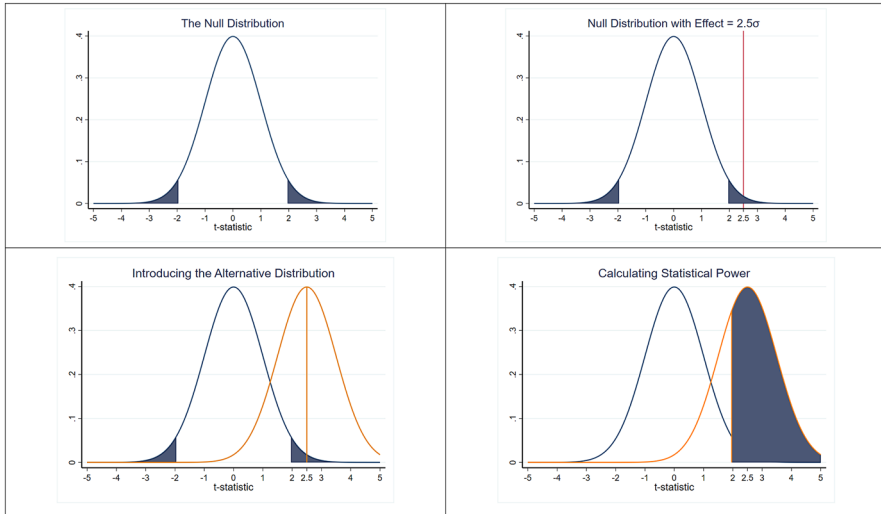


Fig. 1 Illustrating how statistical power is calculated for a two-sided test of statistical significance. The panels in this illustrate how statistical power is calculated for a test of the Null of no difference at the 5% level when the effect is 2.5σ away from the Null value. The rejection region of the test is shown in panel 1, panel 2 shows the effect lands in the rejection region and panel 3 illustrates that the alternative distribution is centered at the effect size. The final panel shows that statistical power is the probability the Null is rejected at the 5% when the effect is 2.5σ away from the Null value

for this default setting that allows for low occurrence of Type 1 error at 5%, but tolerates much greater chance of Type II error at 20%, is to ensure that any discovered new treatment is really effective. After all, switching from the status quo to the new treatment may be costly in many policy areas such as education, health care or social policies. Thus, this design is biased against the new treatment by setting a narrow margin for it to pass, but a much wider margin of falsely concluding that it fails when it did not.

Since one of the primary motivations for undertaking experimental research is to inform theory, not policy with a status quo bias, the rationale for selection of these conventional error probabilities no longer applies.¹² There is no apparent reason to tolerate a substantially greater probability of Type II error than Type I error in a laboratory setting. Further, many economic researchers themselves would not anticipate that a model would hold exactly and would likely be willing to accept a lower

¹² A potential alternative to analyze experimental data is suggested by Manski (2019) who advocates the use of statistical decision theory in place of frequentist hypothesis testing. Motivating this guidance is evidence from Manski and Tetenov (2016) who show that if statistical decision theory is used to determine the sample size, it would be smaller than ones set by conventional statistical power criteria. However, whether using statistical decision theory to analyze experimental data would yield evidence that is informative to theory needs to be explored in future work.

probability of a Type I error.¹³ We next show how the economic theory being tested can inform research design through the estimated standard error in the laboratory.

2.3 Economic theories are quite useful in power calculations

Recall, the between subject design introduced earlier in lab, our sample of subjects, indexed by $i = 1, \dots, n$ to a session that offered either the $D = 1$ or $D = 0$ protocol. For $p = Pr[D = 1]$ such that $1 < pn < n - 1$, we randomly select pn individuals to receive $D = 1$ and the remaining $(1 - p)n$ subjects to receive $D = 0$. Without control variables, the experimental treatment effect for some outcome Y is calculated as $\hat{\gamma}_1 = \frac{1}{pn} \sum_{i \in D=1} Y_i - \frac{1}{(1-p)n} \sum_{i \in D=0} Y_i$. It is well known that when D is binary, $Var(\hat{\gamma}_1) = \frac{1}{n} (\frac{s_1^2}{p} + \frac{s_0^2}{(1-p)})$, where s_1^2 and s_0^2 are sample variances of the outcomes in the $D = 1$ and $D = 0$ groups respectively. Plugging this expression into the minimal detectable effect equation $\gamma_1^{MDE} = (t_{\alpha/2} + t_{(1-\kappa)})\sigma_{\hat{\gamma}_1}$, permits the calculation of the minimum required sample size

$$n^* = \frac{(t_{\alpha/2} + t_{(1-\kappa)})^2 (\frac{s_1^2}{p} + \frac{s_0^2}{(1-p)})}{(\gamma_1^{MDE})^2} \tag{3}$$

If the sample variances are identical between the groups ($s_1^2 = s_0^2$) we can express, $\sigma_{\hat{\gamma}_1} = \sqrt{\frac{\sigma^2}{p*(1-p)*n}}$, where σ^2 is the variance of the residual from Eq. (1). Holding σ^2 and n constant, $\sigma_{\hat{\gamma}_1}$ is minimized when $p = 0.5$, which is when the number of subjects is balanced between the treatment and control groups. Thus, a design that equally splits the sample to $D = 1$ and $D = 0$ ceteris paribus, maximizes statistical power since one could detect a smaller γ_1 for a given level of power (κ) and significance level (α). The variability in the Null and alternative distribution is represented by the standard error that is minimized when $p = 0.5$, and this may explain why the tradition in experimental economics is to assign 50% of subjects to $D = 1$ and the remainder to $D = 0$.

The above calculations ignore the fact that financial cost varies between the two treatments. Cochran (1977) (and subsequently Duflo et al. 2007) shows that if the cost per subject vary across treatments, researchers can optimally choose p to minimize the minimal detectable effect (and hence maximize statistical power) subject to their budget constraint. Solving the Lagrangean indicates that the optimal allocation of subjects between the treatment and control arms is given by the ratio $\frac{p}{1-p} = \sqrt{\frac{c_{D=0}}{c_{D=1}}}$, where $c_{D=0}$ and $c_{D=1}$ is respectively the unit cost per comparison and treated subject.

¹³ For example, (Rodrik 2015, p. 40) states that “Models are never true; but there is truth in models.” Similarly, (Rubinstein 2006, p. 871) writes “When comparing a model to real data, we hope at best to find some evidence that “something” in reality is close to the model’s prediction.”

In other word, the optimal allocation ratio is inversely related to the square root of the relative total costs of each treatment.¹⁴

In many field experiments, the experimenter is ex-ante fully aware of the per subject costs of the different treatments. In lab experiments, the per subject treatment cost depends in large part on the decisions made by the subjects themselves as well as others within the session. However, ex-ante, experimenters set participation fees and often calculate expected earnings in the lab based on the underlying theory. Yet, this information is not used to determine the design of the study and generally an equal number of each treatment is carried out. In the next section, we illustrate how the optimal subject allocation could be calculated using in part these predictions on expected earnings from economic theory in two experimental studies.

2.4 Illustration of statistical power calculation in two experimental studies

We consider two examples: the first price affiliated private value auction of Ham et al. (2005) and the legislative bargaining experiment of Fréchette et al. (2003). In both studies, the subject first received a participation fee of \$5. Otherwise individual compensation differed sharply as one involved subject payment based on cumulative earnings throughout the session, while the other study randomly selected a few periods to determine subject payment.

Ham et al. (2005) conducted sessions with 4-bidders ($D = 1$) and sessions with 6-bidders ($D = 0$) that involved 30 periods. The subjects were paid the total earnings from the auction including a \$7 starting balance as well as earnings from a per-period lottery. The lottery generates an expected earnings of 25 cents per period. Subjects in the 4 bidder sessions are more likely to win more auctions than subjects in the 6 bidder sessions. Moreover, under the symmetric risk-neutral Nash equilibrium bid function (SRNNE), the expected profit to the highest bidder is greater in the 4-bidder sessions relative to the 6-bidder sessions.¹⁵ The ratio of expected costs across treatments per subject is given by

$$\frac{c_{D=1}}{c_{D=0}} = \frac{\text{4-bidder}}{\text{6-bidder}} = \frac{\$5 + (30 * \$0.25) + (\$7 + (\frac{1}{4} * 30 * \$6))}{\$5 + (30 * \$0.25) + (\$7 + (\frac{1}{6} * 30 * \$4))} = 1.6329 \quad (4)$$

Ex ante, the optimal design to maximize statistical power where expected costs are calculated using economic theory is to set $\frac{p}{1-p} = \sqrt{\frac{1}{1+1.6329}} = 0.616$. In other words,

¹⁴ As stated we are making an assumption that is consistent with how the Null hypothesis is stated, that the outcomes for the treatment and control groups have equal means and equal variances. If the variances of the outcomes differ across treatment, the formula has extra terms. The content of this section can easily be modified to incorporate these terms.

¹⁵ Kagel et al. (1987) derive the SRNNE and the high bidder in the 4-bidder sessions is expected to earn \$6 on average compared to \$4 in the 6-bidder sessions. If one restricts attention to values of the private information signal in regions where the SRNNE is approximately linear, expected profits to the highest bidder with parameters from Ham et al. (2005) is approximately $\frac{2i}{n}$; where n is the number of bidders. Ham et al. (2005) restricted their estimation to observations from subjects who were randomly given a signal that fell in this region.

there should be in total 1.63 more subjects in 6-bidder sessions ($D = 0$) relative to 4-bidder sessions ($D = 1$) to detect the smallest treatment effect for a given budget. In Ham et al. (2005), there were a total of 96 subjects. The research interest in Ham et al. (2005) was to estimate the effect of cash balances on bidding behavior. However, if the research interest was in detecting the effect of number of bidders on outcomes and budget constraints are binding, they could have allocated 36 subjects to 4-bidder sessions and 60 subjects to 6-bidder sessions to maximize statistical power.

Similarly, (Fréchette et al. 2003) provide a test of the Baron and Ferejohn (1989) legislative bargaining model on the effect of a closed amendment rule ($D = 1$) versus an open amendment rule ($D = 0$). Subjects took part in sessions in which 15 elections were conducted, at the end of the session 4 elections were randomly selected for payment. Under the parameters selected by Fréchette et al. (2003) any undivided pie would shrink by 20% in the next round of voting. The underlying theory predicts that \$5 expected earnings in each election in the $D = 1$ sessions and \$4.1667 in the $D = 0$ sessions.¹⁶ The ratio of expected costs across treatments per subject is given by

$$\frac{c_{D=1}}{c_{D=0}} = \frac{\text{ClosedRule}}{\text{OpenRule}} = \frac{\$5 + (4 * \$5)}{\$5 + (4 * \$4.1667)} = 1.1538. \quad (5)$$

Using these expected earnings from theory in conjunction with the ratio that chooses the optimal p to maximize statistical power, one would assign subjects to closed versus open rule treatment by setting $\frac{p}{1-p} = \sqrt{\frac{1}{1+1.1538}} = 0.681$. This is an approximate 3:2 ratio for $D = 0$ relative to $D = 1$. That is, to maximize statistical power if we were to recruit 50 subjects, 20 subjects should be assigned to a closed rule session and the remaining 30 subjects should be assigned to an open rule session.

Note, it is common in experiments to observe important quantitative differences between data and the tested theories. For example, in Fréchette et al. (2003) delays were less frequent than theory predicts under the open rule treatment. Similarly, bidders were more aggressive than theory predicts in Ham et al. (2005). If we were to recalculate the optimal ratio based on ex-post experimental earnings, we find in Ham et al. (2005), $\frac{p}{1-p} = \sqrt{\frac{1}{1+1.359}} = 0.651$ (compared to 0.616 calculation from theory) and in Fréchette et al. (2003) $\frac{p}{1-p} = \sqrt{\frac{1}{1+1.056}} = 0.697$ (compared to 0.681 by theoretical prediction). In both cases, we observe that the ratio of subjects across treatment arms remain unbalanced ($\neq 0.5$) and the ex-post ratios both exceed and deviate less from the ex-ante design ratio informed by theory than a even split.

¹⁶ In each election round, a subject chosen at random is selected to propose a division of a \$25 pie across 5 subjects. Theory predicts that under the closed rule, the initial allocation would be approved by a simple majority. Hence, the full size of the pie would be divided according to the proposal made by the randomly chosen subject. In contrast, under the open rule, theory would predict that in each round of an election there is a 50% chance the initial proposal would be amended leading to delays. Since there is a 50% chance that an alternative proposal would take the floor in the next round of the election, the size of the pie would shrink by 20%. The theory predicts that on average there would be a one round delay in the open rule session, in which case, on average \$20 will be divided among the five subjects.

2.5 Power for testing qualitative outcomes

In many experiments, the outcome of interest might be a limited dependent variable. For example, in many bargaining experiments, there is interest in whether a delay in reaching an agreement occurred. For such an outcome, a logit regression model could be deployed to estimate Eqs. (1) and (2), in place of the OLS estimator. Since the logit estimator is nonlinear and involves maximizing a likelihood function, formulas to compute statistical power depend on whether a Wald test, a likelihood ratio test, or score test is used to determine statistical significance.¹⁷ For space considerations here we do not repeat the previous exercise calculating minimum sample size required as well as optimal allocation of subjects into treatment and control. The interested reader can examine Bush (2015)'s Monte Carlo study, which concludes that Shieh's (2000) likelihood ratio test most accurately and consistently achieves the desired level of statistical power.

3 Preanalysis plans can further inform the design of sessions

3.1 Allocating subjects to sessions and possible selection bias

The statistical power calculations presented in the preceding section are used to solve for minimum sample size in total and optimal subject allocation across treatment arms. In this section, we outline how lab experiments could use a pre-analysis plan that contains details on how the experiment is implemented to anticipate potential selection bias issues and remedies. Currently, the implementation of laboratory experiments are often sequential. Lab experimenters frequently send a mass e-mail to their subject pool enlisting sign up and participation in sessions held at specific days or time slots. However, participants in an experimental session on a particular day at a specific time slot can differ based on many factors including those related to course selection, and subjects rarely are allowed to participate in the same experiment twice. Further, the experimenter also could decide immediately prior to the start of the session whether to offer $D = 1$ or $D = 0$, and this decision could potentially be based on which type of subjects have shown up in the session.

Coffman and Niederle (2015) consider sequentiality an advantage since it provides an opportunity to examine the replicability of any finding. We suggest that a form of experimenter session and subject selection bias could arise, if following the arrival of subjects to the lab, the experimenter has the discretion on which treatment

¹⁷ As discussed in Cohen (1988) if additional covariates are included, these formulas also will require the user to ex-ante forecast the squared multiple correlation between the predictor variable and all other variables in the model. Intuitively, when the variable of interest is correlated with other observed confounders, the number of observations needed increases since it is harder to precisely identify the effect of interest. Note that all these complications arise due to the nature of non-linearity with the logit estimator, and could be avoided by using an OLS estimator on a regression model where the outcome variable is a binary variable. Last, Engle (1984) discusses the three testing principles with maximum likelihood methods focusing on the differences in the way the tests look at particular econometric models.

to offer in that particular session,¹⁸ and as discussed in the next subsection who can participate if the session is oversubscribed. If lab experiments are implemented in the manner described above, there is no inherent guarantee that unobserved characteristics of potential on-site participants across treatment arms are balanced.

How subjects are allocated across treatment arms in laboratory experiments contrasts sharply with how it is done in many field experiments in applied economics. In those studies, participating subjects are randomized whether to receive the intervention only after signing up to participate in the study. Randomization is conducted at that point of time in a bid to ensure observed and unobserved characteristics of the participants are balanced across treatment arms.

Lab experiments could use a pre-analysis plan to randomize treatments into specific sessions. This could minimize the threat of selection bias from the experimenter's side by reducing their discretion on what treatment to offer in which session. Further improvement can arise from using block randomization that randomizes sessions within blocks defined by day and time to carry out the session.¹⁹

Alternatively, a lab experimenter could pre-commit to randomize their subject pool into two groups at the beginning of their study. Only members of one group would receive invitations to the $D = 1$ sessions, and the people assigned to the other group would only be eligible to receive invitations to participate in $D = 0$ sessions. This would balance the characteristics of potential subjects prior to signing up to participate in the study.²⁰ Unfortunately, if there are different rates of attendance across sessions, there is no guarantee that one will have random assignment among those that show up (see e. g. Ham and LaLonde 1996).

These two alternatives of allocating subjects to sessions differ from the status quo method by using randomization to remove an experimenter's discretion on which treatment to carry out in a specific session, increasing the chance that both observed and unobserved covariates are balanced across treatment arms. With the experimenter selection biases minimized, the bias from the subjects' self-selective participation remains. Preanalysis plans could ensure appropriate data is collected from either the subject pool or the participating subjects themselves, to both undertake statistical tests of subject selection based on when a session was carried out (e.g. afternoon sessions may disproportionately attract lazier subjects) as well as consider any necessary statistical adjustments to overcome bias arising from the subject's participation decision.²¹ This bias would threaten the validity of the estimated treatment

¹⁸ This action could generate an experimenter session selection bias, which is in the spirit of site selection bias that was initially suggested by Palmer (1993) in the ecology literature.

¹⁹ This would ensure that in each block an equal number of sessions are assigned to each treatment arm, thereby reducing bias and achieving balance relative to a simple randomization scheme.

²⁰ It is important to note that small modifications to the power calculations are needed since the two randomly assigned subgroups can be viewed as strata. Cochran (1977) contains a clear discussion of how the optimal n and p to maximize statistical power are determined with stratified samples.

²¹ As examples, Ham and LaLonde (1996) and Ding and Lehrer (2010) illustrate how to create these adjustments using economic models of the subject's decision to participate respectively in a training program experiment, and education experiment. Related, Casari et al. (2007) address a sample selection problem with both econometric methods and by adjusting economic incentives. They argue that changes in sample design are more likely to be successful than sophisticated methods given the sample sizes in experiments.

effect if there is a lack of balance in the observed and unobserved characteristics of participating subjects in the treatment and control arms.

3.2 Oversubscribed sessions

In practice, it is common to recruit more subjects than needed for each individual session since some subjects that have subscribed fail to show up in person. When a laboratory session is oversubscribed researchers would either offer a financial incentive for some volunteers to leave the study, or use a first-come first-served protocol asking the later arrivals to accept a participation fee and exit the experiment.²²

These usual strategies to handle oversubscription generate selection bias; they also differ from how field experimenters handle this challenge. In the field if a study is oversubscribed, a lottery is often carried out to determine who will receive a treatment.²³ If oversubscription is handled via a voluntary departure induced by a financial incentive, subjects could very well self-select into or out of a study. If first come first served and the late arrivals are students from a class that ended late, the characteristics of the stayers may no longer look like the exiters. Moreover, if the rates of oversubscription differ across the treatments (and sessions), it may no longer be the case that the unobserved characteristics of subjects are balanced across sessions (and treatments). A lottery would give all subjects in an oversubscribed session an equal chance of participating further and can solve this selection problem unless those who show up constitute a random sample.

4 Conclusion

This paper raises three issues that, to the best of our knowledge, have not been discussed in the experimental economics literature. First, what is the appropriate level of the type I and type II errors that a researcher should choose to provide evidence that best informs their intended audience? Second, how many subjects should be allocated to each treatment arm in a study? Third, how should researchers determine which subjects stay in the laboratory when a session is oversubscribed. We suggest that a simple lottery be carried out.

We further argue that there are indeed other benefits from formulating a pre-analysis plan to design the experiment and determine the minimum required sample size, how subjects should be allocated to the control and treatment sessions. We strongly

²² Those who leave the laboratory are often allowed to participate in future sessions of the study. Intuitively, this behavior mimics how airlines oversell their flights and offer individuals credits for future flights to volunteer for and take a later flight.

²³ With data from field experiments, researchers then compare a wait list control group to the experimental group to identify the impact of the given treatment. This strategy assumes that these two groups are comparable since participants were randomly assigned to either the wait list control group or the experimental group. In the laboratory, a no-treatment control group is not used and control sessions need to be carried out, in which there is no guarantee that they will fully populated by members of the wait list.

suggest researchers consider utilizing economic theory to predict the relative cost across the treatment arms when trying to maximize statistical power. We illustrate these strategies and show with data from two laboratory experiments where subject behavior deviated sharply from theory that the use of even split designs as the default design often leads to less statistical power. Finally, we encourage experimenters to adopt new sampling schemes that assign subjects to sessions. These schemes would increase the likelihood that subject characteristics are balanced across the treatment and control arms.

References

- Abreu, D., & Gul, F. (2000). Bargaining and reputation. *Econometrica*, *68*(1), 85–117.
- Baron, D. P., & Ferejohn, J. A. (1989). Bargaining in legislatures. *The American Political Science Review*, *83*(4), 1181–1206.
- Bush, S. A. (2015). Sample size determination for logistic regression: a simulation study. *Communications in Statistics Simulation and Computation*, *44*(2), 360–373.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.
- Casari, M., Ham, J. C., & Kagel, J. H. (2007). Selection bias, demographic effects, and ability effects in common value auction experiments. *American Economic Review*, *97*(4), 1278–1304.
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.
- Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: Wiley.
- Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, *29*(3), 81–98.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
- Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432.
- Ding, W., & Lehrer, S. F. (2011). Experimental estimates of the impacts of class size on test scores: robustness and heterogeneity. *Education Economics*, *19*(3), 229–252.
- Ding, W., & Lehrer, S. F. (2010). Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions. *Review of Economics and Statistics*, *92*(1), 31–42.
- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. In T. Schultz & J. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Amsterdam: Elsevier.
- Embrey, M., Fréchette, G. R., & Lehrer, S. F. (2015). Bargaining and reputation: An experiment on bargaining in the presence of behavioural types. *Review of Economic Studies*, *82*(2), 608–631.
- Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier statistics in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of Econometrics* (Vol. 2, pp. 776–828). Amsterdam: North Holland.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers* (1st ed.). Oliver and Boyd: Edinburgh.
- Ford, I., Norrie, J., & Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, *14*(8), 735–746.
- Fréchette, G. R., Kagel, J. H., & Lehrer, S. F. (2003). Bargaining in legislatures: an experimental investigation of open versus closed amendment rules. *American Political Science Review*, *97*(2), 221–232.
- Ham, J. C., Kagel, J. H., & Lehrer, S. F. (2005). Randomization, endogeneity and laboratory experiments: The role of cash balances in private value auctions. *Journal of Econometrics*, *125*(1–2), 175–205.
- Ham, J. C., & LaLonde, R. (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica*, *64*(1), 175–205.
- Heonig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*(1), 19–24.

- Hernandez, A. V., Steyerberg, E. W., & Habbema, D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, *57*(5), 454–460.
- Kagel, J. H., Harstad, R. M., & Levin, D. (1987). Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica*, *55*(4), 1275–1304.
- List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439–457.
- List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.
- Maniadis, Z., Tufano, F., & List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal*, *127*(605), F209–F235.
- Manski, C. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician*, *73*(s1), 296–304.
- Manski, C., & Tetenov, A. (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences*, *113*(38), 10518–10523.
- Nikiforakis, N., & Slonim, R. (2015). Editors preface: Statistics, replications and null results. *Journal of the Economic Science Association*, *1*(2), 127–131.
- Palmer, M. W. (1993). Potential biases in site and species selection for ecological monitoring. *Environmental Monitoring and Assessment*, *26*, 277–282.
- Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, *59*(2), 227–240.
- Rodrik, D. (2015). *Economics Rules: The Rights and Wrongs of The Dismal Science*. New York: W.W. Norton.
- Roth, A. E. (1986). Laboratory experimentation in economics. *Economics and Philosophy*, *2*, 245–273.
- Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, *56*(4), 1192–1196.
- Slonim, R., Wang, C., Garbarino, E., & Merrett, D. (2013). Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior and Organization*, *90*(1), 43–70.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.