Measuring the Effects of Experimental Costs on Sample Sizes

Jason M.T. Roos*

5 May 2017

Abstract

This study uses archival data from a social science behavioral lab and standard quantitative tools from consumer research to generate the first empirical measure of how experimental costs affect sample sizes. Researchers at this lab are sensitive to higher experimental costs, and more specifically, to the amount of money or course credit needed to reimburse study participants. Model estimates are used to assess counterfactual lab policies which, by lowering researchers' costs, incentivize larger samples. At this lab, a 50% subsidy on cash reimbursements (normally paid from individual research budgets) would have been expected to increase sample sizes by 35% among studies originally run with 50–200 participants. The method used for this analysis provides a flexible approach to studying experimenter behavior and research policies at other labs, and more generally, demonstrates the value of considering researchers' decisions in a consumption framework as part of efforts to improve scientific outcomes.

Keywords. Research policy, structural models, Bayesian estimation, experimental design

*Associate Professor, Rotterdam School of Management and ERIM, Erasmus University, PO Box 1738, 3000 DR Rotterdam, Netherlands, +31 10 408 2527, roos@rsm.nl. Thanks to Sara Rafael Almeida for help obtaining the data used for this study, as well as Stefano Puntoni, Bram Van den Bergh, Mirjam Tuk, Ale Smidts, Carl Mela, Ron Shachar, Leif Nelson, Gabriele Paolacci, Steven Sweldens, Alina Ferecatu, Amit Bhattacharjee, Dan Schley, Begüm Şener, Martina Pocchiari, and seminar participants at RSM, the Erasmus/Tilburg JDM Camp, the Bayesian Econometric Forecasting and Policy Analysis Workshop, and the University of Groningen. This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

1 Introduction

Scientists have finite resources to carry out their work. In the context of experimental or survey-based research, this constraint means researchers typically choose sample sizes in the face of two competing incentives. One is the incentive to gain knowledge about a population of interest with maximum precision. Holding fixed the design of the experiment or survey, precision can be increased by collecting a larger sample. But competing against this incentive is another: the desire to minimize the time and expense needed to obtain that knowledge. Holding fixed a study's design, such costs can be decreased by collecting a smaller sample. Hence, in most situations, a study's sample size reflects a trade-off, at some level, between the higher precision of a bigger sample and the lower cost of a smaller one (Blattberg 1979; Cohen 1992b; Allison et al. 1997; Gelman and Carlin 2014).

A conflict can occur if these competing incentives lead researchers to make choices that are misaligned with the goals of their stakeholders, whether they be funding agencies, journal editors, or simply the broader scientific community (Dasgupta and David 1994). More specifically, researchers might choose to conduct experiments with samples sizes which they deem sufficient to answer a particular question, but which others consider inadequate or "underpowered" (i.e., having too high a probability of false negative or Type II error). Consider, for example, a planned two-cell experiment with 50 participants per condition, each of whom will be paid $5. Adding five participants to each cell would cost the researcher an additional $50, but also raise the study's power. Is the higher cost worth the higher chance of detecting a true effect? A researcher with limited funds, but expecting only a small increase in power, might not think so, whereas others might (strongly) disagree.

Low-powered studies have been a persistent problem in the social sciences for more than fifty years. In a meta-analysis of results published in the *Journal of Abnormal Psychology,* Cohen (1962) estimated (post hoc) median power to be .17, .46, and .89 for small, medium, and large effects—meaning an experiment measuring a *true* difference in means of .15 standard deviations had just a

1 in 6 chance of rejecting the null (at $\alpha < .05$). This surprising result spurred the development of tools for conducting a priori "power analysis" (Cohen 1969) and helped to raise awareness about the benefits of bigger samples. But in spite of these positive developments, subsequent replications of Cohen's analysis show that efforts to get experimental researchers to run higher-powered studies—including better statistics training, free power calculation software, and innumerable editorials—have historically had very little impact on their behavior (Sedlmeier and Gigerenzer 1989; Maxwell 2004; Shen et al. 2011; Marszalek et al. 2011).

There are numerous ways to increase experimental power, many of which have been encoded as best practices in the area of experimental design (Allison et al. 1997; McClelland 2000; Abraham and Russell 2008; Button et al. 2013). The focus of this paper, however, is on achieving higher experimental power via larger samples *after* these design decisions have been made. There is increasing recognition of the need for bigger samples, especially in light of recent concerns over replicability in the social sciences (Cohen 1992b; Maxwell 2004; Ioannidis 2005; Shen et al. 2011; Simmons et al. 2011; Bakker et al. 2012; Schimmack 2012; Asendorpf et al. 2013; Button et al. 2013; Miguel et al. 2014; Simmons 2014; Meyer 2015; however see Baumeister 2016 for counterarguments). Moreover, as efforts to control false positive or Type I errors (which to date have received greater attention than false negative rates) gain ground, the need for larger samples grows more pressing (Fiedler et al. 2012; Simmons et al. 2013).

From the perspective of funding institutions and the broader scientific community, running experiments with little chance of measuring the hypothesized phenomenon is wasteful in terms of the money used to compensate participants, as well as researchers' time (and not just that of the experimenters themselves, but also those who rely on published results, which are more likely to be erroneous; Button et al. 2013). This paper seeks to provide an answer to the question of how an institution might incentivize a researcher who would otherwise run a study with a smaller sample (as measured by some external standard), to instead run that same study with a bigger sample.

Unlike previous work which has only considered researchers in their role as producers of new knowledge (Stephan 1996; Dasgupta and David 1994), this paper takes a different approach by

considering researchers in their role as *consumers* of the information generated by their study participants. Under this approach, one can think of each additional study observation "consumed" by the researcher as increasing the expected payoff from running the study, but at the same time increasing its total cost. The researcher deciding whether to add 10 more participants at a cost of $50 is seen as weighing these higher costs against an expected change in reward due to running a higher-powered study.

Framing the experimenter's choice of sample size this way—as a trade-off between better science and higher costs—immediately highlights two broad strategies that could potentially increase researchers' willingness to collect larger samples. One strategy would be to increase the benefits of running experiments with larger samples, but this approach is problematic because the rewards from obtaining publishable results are extremely high (Nosek et al. 2012; Button et al. 2013; Miguel et al. 2014), and researchers already face strong incentives favoring large samples (Maxwell 2004). Indeed, the strong incentive researchers have to yield publishable results is frequently blamed as a root cause for the prevalence of small samples and other practices leading to Type I errors (Ioannidis 2012b; Button et al. 2013). The other strategy is to decrease the costs associated with collecting bigger samples. Indeed, if we think of researchers as consumers of experimental samples who incur higher costs (in terms of money or time and effort) whenever they collect data, then cost-reducing policies implemented at the institutional level would seem to be a straightforward route to incentivizing bigger samples.

The primary obstacle to designing such policies, however, is a lack of empirical research measuring the relationship between experimental costs and sample sizes. Although previous studies have explicitly addressed the trade-off between the costs and benefits of conducting experiments, this work has been either theoretical or prescriptive (Blattberg 1979; Ginter et al. 1981; Sawyer and Ball 1981; Chatterjee et al. 1988; Cohen 1992a; Allison et al. 1997; Moscarini and Smith 2002; Winkens et al. 2006). Missing from this literature are empirical studies measuring how much experimental costs actually influence sample sizes. Such measurements are needed to assess whether the expected gains from new policies would be likely to offset their implementation costs.

The goals of this paper, therefore, are the following: 1) to provide the first empirical measure of how researcher costs affect sample sizes and experimental power at a behavioral laboratory, 2) to assess for this particular lab how sample sizes might have increased under counterfactual institutional policies that would have lowered researchers' experimental costs, and 3) to provide a general methodology, based on standard tools from consumer research, that can be used to study researcher behavior and evaluate cost-lowering policies at other labs.

With these objectives in mind, an empirical model of experimental sample sizes is estimated from data describing experiments conducted at the same behavioral laboratory by a diverse group of social scientists. These data are unique to the literature and provide a much-needed view into how behavioral researchers conduct their work on a day-to-day basis.

Previous studies have only considered sample sizes as reported in published journal articles (Cohen 1962; Sawyer and Ball 1981; Sedlmeier and Gigerenzer 1989), which are subject to publication bias (the so-called "file drawer" problem; Rosenthal 1979). This study, however, uses data obtained directly from the lab's participant scheduling system, and thus includes information about both published and unpublished experiments. Inclusion of the latter group is a crucial step in obtaining an accurate measurement of how costs affect sample sizes, and a novel contribution of this paper.

The empirical model treats observed sample sizes as the outcome of a choice process that weighs the expected benefit from running the experiment against two types of cost: 1) the amount paid to each participant, and 2) the time spent collecting data. Researchers in this model pay to consume the information provided by study participants, and hence their chosen sample sizes reveal their preference for larger samples and sensitivity to higher costs.

This modeling approach borrows heavily from standard quantitative techniques for understanding consumers in other choice settings, in which the choice model is a simple approximation of a more complex decision process, but one that includes many of the factors that might be important to the decision maker (and thus useful for the policy maker to understand). Accordingly, the purpose of this exercise is not to show that experimental costs matter—we already know this. Rather,

the goal is to measure *how much* these costs affect sample sizes so that we can design better institutional policies. Estimates show, for example, that the disutility from reimbursing participants at this lab has a far greater effect on sample sizes than the time required to collect the observations. This suggests that policies aimed at reducing the amount researchers must reimburse participants (e.g. via subsidies) would be more effective at this lab than policies aimed at reducing the number of days needed to collect a bigger sample.

Using the model estimates, cost-lowering policies are simulated in order to understand how these might have affected researchers' chosen sample sizes. These simulations confirm that subsidizing participant reimbursements has the potential to meaningfully reduce the disincentive against obtaining a bigger sample. Even more importantly, these simulations generate estimates for the magnitude of these gains. A 50% subsidy in cash reimbursements, for example, would be expected to increase sample sizes on average by as much as 35% at this lab. Finally, survey data reported in Gervais et al. (2015) are used to estimate how these increases in sample size might translate to increases in experimental power, providing an alternative metric for understanding the expected impact of the counterfactual policies.

2 How Costs Can Affect Sample Size and Experimental Power

Before presenting the data, analysis procedure, and results, it is helpful to illustrate how costs can affect sample size (and by extension, experimental power)—even when the researcher is well trained and highly scrupulous—by comparing two stylized approaches to choosing a hypothetical sample size for a simple, preregistered, confirmatory experiment.[1] The first approach is based on Cohen's (1969; 1992) suggestion to power studies at no less than .80; the second on expected utility maximization (Sawyer and Ball 1981; Gelman and Carlin 2014). Details relevant to both

[1] Although the assumption of a scrupulous and well-trained researcher is not necessary for the arguments advanced in this section, it highlights an important point: The goals of the institution and researcher can be misaligned for reasons that have nothing to do with the related, but different problem of Type I, or false positive, error control. Consequently, researchers who are experts in statistics, who always pre-register their experiments, and who never "*p*-hack" (Simonsohn et al. 2014), may nevertheless make well-informed decisions about experimental power and sample size that seem fine to them, but look like bad science to interested and well-meaning observers (Button et al. 2013).

Table 1 Prior Beliefs, Results, and Payoffs from a Hypothetical Experiment

PRIOR BELIEFS		*EXPERIMENT*		*EXPECTED PAYOFFS*		
State	*Probability*	*Result*	*Probability*	*Amount*	*Probability*	*Description*
H	q	B	$1-\beta_n$	$u(B,H)$	$q(1-\beta_n)$	Null rejected
H	q	A	β_n	$u(A,H)$	$q\beta_n$	Type II error
L	$1-q$	A	$1-\alpha_n$	$u(A,L)$	$(1-q)(1-\alpha_n)$	Null not rejected
L	$1-q$	B	α_n	$u(B,L)$	$(1-q)\alpha_n$	Type I error

approaches are discussed first.

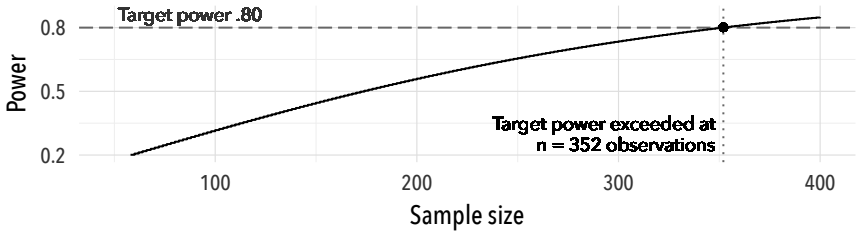
2.1 A Hypothetical Experiment

A researcher must choose a sample size, n, for a two condition (between subjects) test of a simple hypothesis. Each participant is reimbursed at a rate of p (although units are irrelevant here, one can think of costs and payoffs in money terms). H denotes the state of the world in which the hypothesized effect truly exists (i.e., the alternative hypothesis is correct), and L the state of the world in which it does not (i.e., the null is correct). Initially, the researcher believes the effect is real (i.e., we are in state H) with probability q, and if it exists, has a standardized effect size $\bar{d} = .3$. The experiment produces one of two outcomes, depending on whether a t-test rejects (result B) or fails to reject (result A) the null.

With two experimental results and two potential states of the world, there are four possible outcomes, as shown in Table 1. Each of these outcomes is associated with an expected payoff, $u(\cdot,\cdot)$. For example, the payoff from correctly detecting a true effect, $u(B,H)$, might reflect the anticipated joy and increased future earnings that result from having an interesting finding to write up and publish in an academic journal.

The expected Type I and II error rates are denoted α_n and β_n. Theoretically both probabilities depend on the chosen sample size, but in most applied settings α_n is set to .05 for any sample size n, causing any increases in sample size to impact β_n alone. Given the researcher's hypothesis, prior beliefs, expected payoffs, and costs, the next step is to choose a sample size n for the experiment.

Figure 1 Power Analysis Approach to Sample Size Selection for a Hypothetical Experiment



Notes. Power calculations assume a *t*-test of the difference in sample means with expected effect size $\bar{d} = .3$.

2.2 The Power Analysis Approach

The most widely endorsed normative approach to choosing sample size is to first conduct an ex ante power analysis, then choose the smallest n yielding Type I and II error rates less than .05 and .20 respectively—i.e., experimental power, $1 - \beta_n$, should be no less than .80 (Cohen 1969, 1992a, 1992b; Lenth 2001; VanVoorhis and Morgan 2007; Maxwell et al. 2008; Simmons et al. 2013). Figure 1 shows how in this example, the target error rates are satisfied with at least $n = 352$ (176 observations per condition).

Cohen (1969) first proposed setting n to achieve a target power of .80, and although his suggestion is widely known, the rationale behind the choice of .80 is not:

> In scientific research, it is typically more serious to make a false positive claim (Type I error) than a false negative one (Type II error). Because the implicit convention for significance is $\alpha = .05$, the use of the .80 convention for desired power (hence, $\beta = .20$) makes the Type II error 4 times as likely as the Type I error, an arbitrary but reasonable reflection of their relative importance (Cohen 1992b, p. 100).

Cohen intended .80 to be a default, not a strict requirement. Nevertheless, it has become a normative standard in spite of being, in his words, "arbitrary."

But more important than the arbitrariness of .80 is the following: Under the power analysis approach, sample size depends on the expected effect size and target error rates. Conspicuously absent from this decision calculus are: 1) the expected payoffs from the experiment's outcomes,

2) the researcher's prior belief about the veracity of the hypothesis, and 3) the cost of collecting data. The importance of these omitted factors provides a simple (albeit partial) explanation for researchers' collective failure to rely exclusively on ex ante power analysis when choosing sample sizes, as illustrated next.

2.3 The Expected Utility Approach

Given the researcher's prior beliefs and expected utilities, the expected payoff from the experiment can be written as a function of the sample size n (Moscarini and Smith 2002):

$$V(n) = \underbrace{q\Big[\underbrace{(1-\beta_n)\,u(B,H)}_{\text{Null rejected}} + \underbrace{\beta_n u(A,H)}_{\text{Type II error}}\Big]}_{\text{Effect exists } (H)} + \underbrace{(1-q)\Big[\underbrace{(1-\alpha_n)\,u(A,L)}_{\text{Null not rejected}} + \underbrace{\alpha_n u(B,L)}_{\text{Type I error}}\Big]}_{\text{No effect } (L)} \quad (1)$$

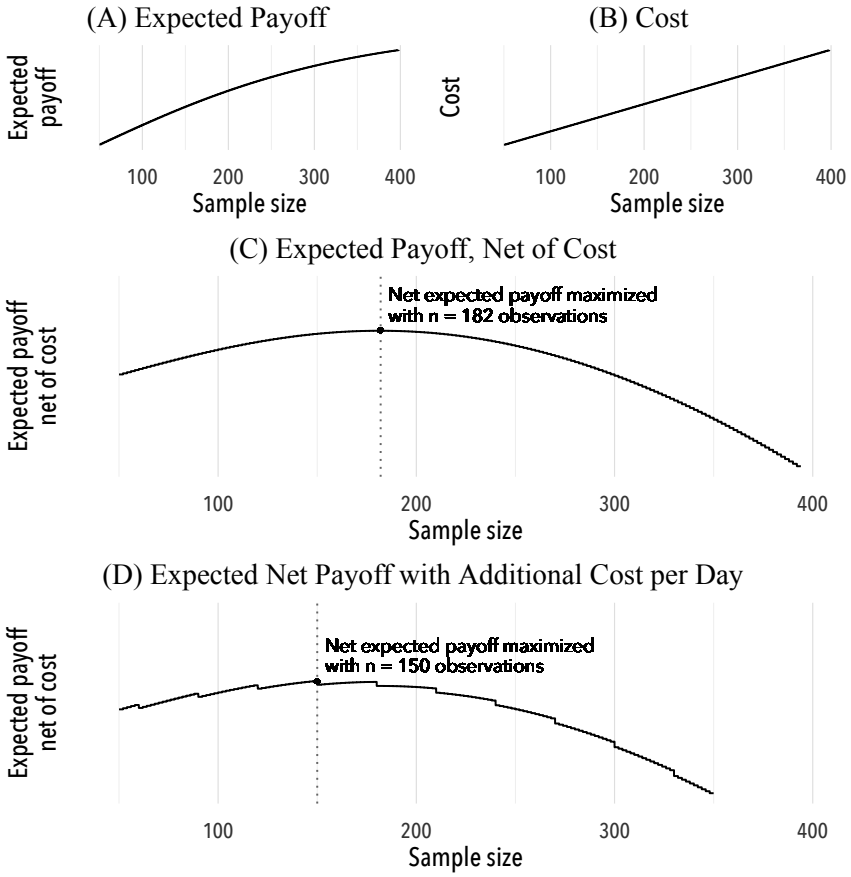
The payoff function $V(n)$ is a weighted average of the expected utilities (the u's), with the weights determined by the researcher's prior belief about the hypothesis (q) and expected error rates (α_n and β_n). The expected error rates are themselves functions of the chosen statistical test, expected effect size $\bar{d}$, and sample size n.

Because larger samples reduce the Type II error rate, the expected payoff is increasing in n, $V(n+1) > V(n)$. But as n gets very large, the improvement grows smaller and smaller, as shown in Figure 2A. In other words, there are diminishing marginal returns from bigger samples, hence $V(n+1) - V(n) < V(n) - V(n-1)$. Nevertheless, if obtaining observations were costless, the researcher would include as many participants as possible in the experiment, since each additional observation increases the experiments's expected payoff.

Obtaining a massive sample, however, is *not* costless. Rather, each participant receives a payment of p, which leads to a total experimental costs that increases linearly in n, as shown in Figure 2B.

Because larger samples increase the expected payoff at an ever-diminishing rate, but increase experimental costs at a constant rate, there is always a sample size n at which the marginal benefit from collecting the $n + 1^{\text{th}}$ observation is less than its cost (Blattberg 1979; Moscarini and Smith

Figure 2 Expected Payoff, Costs, and Payoff Net of Costs, from a Hypothetical Experiment



Notes. The expected utilities in this example are $u(B, H) = 1{,}000p$, $u(A, H) = 100p$, $u(A, L) = 200p$, and $u(B, L) = 0$; and the prior belief is $q = .5$. Because the researcher in this example does not *p*-hack, true negatives generate greater utility than false positives (hence $u(A, L) > u(A, H)$). Skeptical readers should note however that the expected utilities, prior beliefs, and effect sizes chosen for this illustration have no bearing on the main qualitative result, and for this reason, the figures do not show numerical units along the *y*-axes.

2002). Hence, the researcher's choice problem can be framed as one of expected utility maximization:

$$\text{choose } n > 0 \text{ to maximize} \quad V(n) - C(n), \tag{2}$$

where $V(n) - C(n)$ equals the experiment's expected payoff, net of costs, given a sample size of n. As Figure 2C shows, net expected payoff is highest in this toy example with a sample of $n = 182$, producing a study powered at .52.

Although this hypothetical experiment is powered lower than the normative standard of .80, it is—to the researcher at least—optimal. Indeed, from the researcher's perspective, it would be difficult to justify powering this experiment at .80 given the high marginal cost and low marginal benefit from any additional observations. Importantly, this result arises even though the researcher preregistered the experiment, conducted a power analysis (recall the researcher calculates β_n in Equation (1)), and is (by assumption here) incapable of *p*-hacking.

Finally, if other costs bear on the researcher's decision, the sample size may be even lower. For example, Figure 2D shows what happens if a capacity constraint limits the experiment to 30 observations per day, and the experimenter incurs $2p$ of disutility for each day collecting data. In this case, there are values of n at which an additional observation would require one more day in the lab, producing a sharp decline in net expected payoff. In the example in Figure 2D, the chosen sample size is $n = 150$, and the experiment is powered at .45.

2.4 Discussion

The purpose of this illustration is to show the simple but powerful influence experimental costs exert over sample sizes. These contrasting choice models make a simple point: Cost-sensitive researchers—even statistical experts who never cheat—face incentives leading to decisions that are optimal from their perspective, but potentially suboptimal in the eyes of key stakeholders.

In contrast to other settings where we study consumption decisions, we hold researchers to a normative standard in which they are not cost-sensitive, hence the power analysis approach is a widely accepted normative model of sample size selection. But considering what we know about

consumer choice in other domains, it should come as no surprise that experiments powered greater than .80 are the exception and not the rule.

Of course, one may re-cast the power approach as a special case of expected utility maximization, wherein the cost function $C(n)$ is equal to infinity for every value of n except the smallest one achieving $\beta_n < .20$. But the absurdity of this formulation only further illustrates how asking researchers to power experiments at .80, without simultaneously addressing their incentive to minimize costs, will not likely change their behavior.

It should be stressed that the value of specifying expected payoffs as in Equation (1) comes not from the equation's formalism, but rather from its explicit consideration of researchers' prior beliefs and expected utilities. The point here is not to propose, either descriptively or normatively, that researchers in the real world actually quantify their prior expectations and utilities as precise numerical values. More likely, they follow the decision process in Equation (2) using a fairly accurate accounting of $C(n)$, and a heuristic approximation to $V(n)$ (Maxwell 2004). The approximation to $V(n)$ used for the empirical analysis has exactly this essential character.

But even though this model is a simplified approximation of a more complex choice process, it still retains a great deal of flexibility. For example, many researchers rely on "rules of thumb" when choosing sample sizes. However, as many have noted, researchers have many such rules to choose from (Sawyer and Ball 1981; VanVoorhis and Morgan 2007; Maxwell et al. 2008). Hence a researcher whose rule is "20 observations per cell" instead of "30 observations per cell," or "50% power" instead of "80% power," has still chosen a level of statistical efficiency in the face of experimental costs, and can therefore be described by Equation (2) (e.g., by restricting n to values permitted by the myriad rules of thumb available, $n \in m \times \{10, 15, 20, 30, 50\}$, with m equal to the number of experimental conditions).

Although this paper is motivated by issues of Type II error control, Equation (1) can also speak to how changes in Type I control affect sample sizes. First, because true false positive rates are usually much greater than the nominal .05 (Simmons et al. 2011), improvements in research practices (e.g., study pre-registration) will yield more stringent significance levels, which in turn will

correspond with lower power. Hence, as efforts to control false positive rates make further inroads, researchers will need to work with bigger samples (Fiedler et al. 2012), but might not immediately recognize the need to do so. Second, decreasing the payoff from false positive results (e.g., through some sort of punishment) corresponds with a lower value of $u(B, L)$, thus shifting the curve in Figure 2C downward. This downward shift lowers the total expected value of the experiment—perhaps making it less likely to be conducted in the first place—but leads to exactly the same optimal sample size.

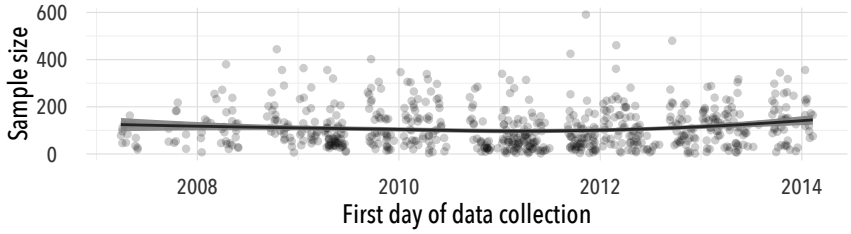
In order to design policies incentivizing the use of larger samples, it is not enough to simply know that higher costs can lead to smaller samples. Rather, we need to empirically measure the magnitude of this relationship. The next two sections present the data and model used to measure this relationship at one particular lab.

3 Archival Lab Data

The data used for this study were obtained from the participant management system at the Erasmus Behavioral Lab, a joint research facility operated by the Institute of Psychology and the Erasmus Research Institute of Management at Erasmus University Rotterdam in the Netherlands. The data describe all experiments conducted at the lab between the inception of a credit-reimbursed participant pool on March 30, 2007 (a paid participant pool was introduced in October 2008), and February 21, 2014. These experiments were performed by a diverse group of social science researchers at all academic ranks (including graduate students), most of whom were affiliated with one of three university divisions: the Faculty of Social Science, the Erasmus School of Economics, and the Rotterdam School of Management (RSM). The most active users of the lab were affiliated with RSM (and in particular, the Department of Marketing Management).

The data are organized according to whether participants were reimbursed with course credit ("credit pool") or money ("paid pool"), and include (among other variables) each experiment's title and description, expected duration, amount of money or course credit paid to participants, and a list of researchers associated with the study. Titles, descriptions, and experimenter lists identify and

Figure 3 Sample Sizes over Time



Notes. Each point represents an experiment. Empty vertical regions correspond with the summer and winter holidays. The LOESS regression line (with 95% CI in grey) shows typical sample sizes have not changed much over time.

link a subset of experiments that included participants drawn from both pools.

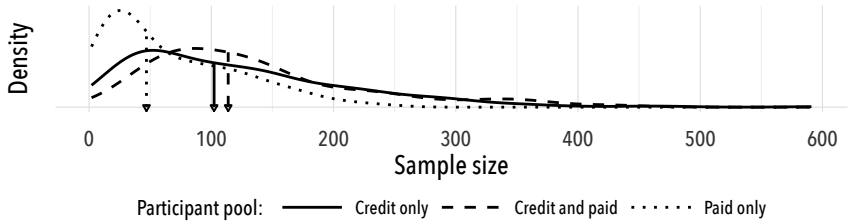
3.1 Experiments

For each experiment, the data indicate the exact timing of each participant's involvement in the study. From this, the total sample size, number of days in the lab, number and timing of observations on the last day of data collection, and other statistics are calculated. A small number of observations describe experiments for which data collection took place outside the behavioral lab (e.g., at the nearby Erasmus Medical Center). As this analysis seeks in part to understand how time in the lab influences sample sizes, these experiments are excluded from the analysis (this decision preceded any statistical analysis of the data; please see Appendix A for further details regarding data preparation). The data used for estimation describe 683 experiments (63% in the credit pool, 26% in the paid pool, and 11% in both) associated with 134 researchers.

Figure 3 illustrates how activity within the lab varied over time. Although there is strong evidence of seasonal variation in lab use due to summer and winter holidays, there is no clear indication that average sample sizes have changed much over the 7 years recorded in the archival data.

Sample sizes do differ significantly between experiments conducted exclusively in the paid and credit pools, as shown in Figure 4. The median sample size across all studies is 89, but 102 for credit pool-only studies and 47 for paid pool-only studies (114 for studies using both pools). This diversity suggests there may be systematic differences in the types of experiments conducted in the two participant pools. For example, paid studies might differ from others due to the availability of

Figure 4 Distribution of Sample Size by Participant Pool



Notes. Contours represent kernel-smoothed densities of sample sizes in the three participant pools. Vertical lines mark median sample sizes.

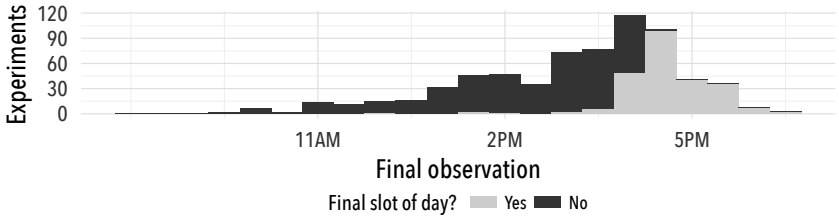
money-incentive manipulations, non-student or more attentive respondents, a higher prevalence of pre-tests, etc.

3.2 Researchers

Experiments vary systematically depending on the number of researchers involved. The median sample size is 76 among the 71% of experiments conducted by one researcher, but 117 among the 25% associated with 2–3, and 155 among the 4% associated with 4–7. The nature of these collaborations is not observed. In many cases, one or more of the collaborators is a graduate student (possibly a research assistant); in other cases, collaborators have combined multiple unrelated experiments into a single session. Both suggest teams of researchers might have pooled resources, possibly in order to obtain bigger samples (Stephan 1996).

The only data describing individual researchers are their email addresses. In many cases these identify affiliated institutions or whether researchers are graduate students or faculty. A small number of groups are defined, and each researcher is assigned to one. These groupings provide a source of observed heterogeneity, and their inclusion in the model improves its efficiency. But because understanding differences in sample sizes across groups of researchers is beyond the scope of this paper, and moreover, because presenting results related to these data could be harmful to specific researchers, these results are not reported numerically.

Figure 5 Distribution of Timing of Final Observations by Time of Day



Notes. Stacked histogram showing start times of final observations in half hour increments. Shading indicates whether observations occurred at the last possible time of the day (based on time required to collect one observation).

3.3 Cost Details

For each experiment, data describing two types of cost are available. First there is the amount paid to each participant. Participant reimbursement is closely related to the time needed to collect each observation. The norm in this lab is to pay participants at a marginal rate of 1 course credit or €5 per half hour. 99% of studies that used course credit to compensate participants followed this norm exactly. Among the studies which compensated participants with cash, 74% adhered to this rule exactly, 12% paid less, and 14% paid more.

Although the notion that costs affect sample sizes is well accepted, it nevertheless remains an assumption of this model. Appendix B describes tests of conditional independence (Pearl 2009) which show that variation within the archival data is consistent with this assumption (and highly unlikely to occur otherwise).

The second cost is due to time spent in the lab, including: 1) the number of days of data collection, 2) the timing of the last observations on the final day of data collection, 3) the time needed to collect each observation, and 4) the number of experiments sharing the lab each day. On the final day of data collection, participants were frequently scheduled through the end of the day, as shown in Figure 5. Hence, a study represented by the lighter bars in Figure 5 could not have increased its sample size from n to $n + 1$ without extending data collection into the next day. If the cost of that extra day in the lab were high enough, it might have offset the gain from obtaining the $n + 1^{th}$ observation, leading to the pattern shown in Figure 5 (if the cost of that extra day were nil, then a more uniform distribution of end-times would be expected). The median study took 30 minutes

to administer and occupied the the lab on 4 separate days, sharing it with an average of 3 other experiments each day. If spending time in the lab is costly to researchers (e.g., due to social pressure not to overuse shared resources; Gneezy et al. 2014), then this cost should be highest among researchers using the lab at the busiest times.

Researchers compensate participants in the paid pool using their individual research accounts, but they have no analogous budget for reimbursing students with course credit. Moreover, because there is an institutional guarantee of participant reimbursement, but no check to ensure the researcher's (money) budget can cover the expense, researchers always have the option of working with a slightly larger sample size. These details bear directly on the definition of the model likelihood (see Appendix C for further details).

3.4 Discussion

The data used in this study are unique in the literature on sample size and experimental power. They describe a wide range of experiments, including "successful" studies, "unsuccessful" studies, pre-tests, tests of major hypotheses, and everything in between. It doesn't matter whether the researchers analyzed their observations with state-of-the-art statistics or rudimentary procedures; whether they pre-registered their studies or *p*-hacked them—every study conducted at this lab is reported in the archive. Whatever the researchers did after the data were collected, however, is not.

This study thus measures how experimental costs affected sample sizes at this lab, regardless of what the researchers subsequently did with their data. For this reason, the empirical results neither suffer from publication bias, nor depend on assumptions about the researchers' statistical expertise or tendency to engage in questionable research practices. Indeed, the main assumption for conduct is that researchers behaved consistently over time. On the other hand, the data lack the necessary information for modeling researchers' choices of which studies to run, as well as participants' choices of which studies to enroll in. This absence of data prohibits quantifying the extent to which the cost interventions considered later might increase the number of experiments conducted.

Finally, it should be emphasized that the archival data describe activity at just one lab, limiting the generalizability of the empirical results. This particular lab has a high capacity by comparison to labs at similar institutions, occupying over 6,000 sq. ft. and averaging more than 275 observations on the 10 busiest days. The time cost of data collection is probably less salient at this lab than most, whereas researchers at institutions with smaller facilities but greater research budgets would find the opposite to be true.

4 Empirical Model of Sample Size Choice

This section describes an empirical model corresponding with the choice problem defined by Equation (2). That is, for each experiment j, one or more researchers chooses a sample size to maximize the expected benefits from the experiment, net of costs.[2] To clarify the exposition, the model is first presented in a simplified context: that of a single researcher choosing sample sizes for experiments that are very similar in their payoffs and design (e.g., multiple conceptual replications of the same phenomenon). After presenting the researchers' cost and payoff functions in this simplified setting, the model is then augmented to accommodate heterogeneity in the types of experiments conducted (as well as in the researchers who run them). This section concludes with a brief discussion of the model likelihood function; other estimation details are included in Appendix C.

4.1 Experimental Costs

The researcher incurs a cost for running experiment j with n participants. This cost is decomposed into three parts: 1) the money or course credit paid to each participant, 2) the disutility from each day of lab use, and 3) a one-time setup cost for the entire experiment.

$$C_j(n) = \underbrace{p_j n}_{\text{Participants}} + \underbrace{\lambda\left[D_j(n)\right]^{\delta}}_{\text{Lab time}} + \underbrace{F_j}_{\text{Setup}}, \qquad \lambda > 0, \quad \delta > 0 \tag{3}$$

[2] This choice is conditioned on having already decided to conduct the experiment in one of the subject pools and settled on a design, including the time needed to complete the task, and by extension, the amount to be reimbursed to each participant (recall about 70%/99% of experiments using the paid/credit pool paid exactly €5/1 credit per half hour).

The first term on the right-hand side of Equation (3), $p_j n$, indicates the total amount paid to participants. For experiments using participants from both the paid and credit pools, p_j is a weighted average of the cash (euros_j) or course credit (credits_j) remitted to each person (for studies run exclusively in one pool, either euros_j or credits_j is zero).

$$p_j \equiv \text{euros}_j + \alpha\,\text{credits}_j, \qquad \alpha > 0 \tag{4}$$

The parameter α translates credit payments to a money scale.

The second term on the right-hand side of Equation (3), $\lambda\left[D_j(n)\right]^{\delta}$, represents the researcher's disutility from using the lab for $D_j(n)$ days. The number of days required for the experiment, $D_j(n)$, is an increasing, stepwise function of the experiment's sample size, and is observed in the data. The parameter δ permits either increasing ($\delta > 1$), decreasing ($0 < \delta < 1$), or constant ($\delta = 0$) marginal costs from each additional day, and the parameter λ translates disutility from time in the lab to the same scale as the participant reimbursement costs.

The final term in Equation (3), F_j, represents any setup costs. Because the archival data do not contain information about setup costs, the value of F_j cannot be estimated (see Appendix C). Parameter estimates and the results of the counterfactual analysis should therefore be interpreted in the context of the studies that were actually conducted in the lab (and not in the context of all possible studies, including hypothetical studies that could have been run, but were not).

4.2 Expected Payoffs

Because the data do not include information describing researchers' states of mind (e.g., prior beliefs, expected utilities, etc.) when they planned their experiments, the expected payoff in the model, $V(n)$, is based on the reduced-form approximation to Equation (1) developed by Moscarini and Smith (2002). This approximation is derived from the general statistical properties of significance tests, and as such, overcomes the obstacle of not observing researchers' expected payoffs, prior beliefs, or expected Type I/II error rates in the data. The expected payoff when running experiment

j with a sample size of n is:

$$\widehat{V}_j(n) = V_j^* - \kappa \frac{\rho^n}{\sqrt{n}} e^{\epsilon_j}, \qquad \rho \in (0,1), \quad \kappa > 0 \tag{5}$$

The first term on the right-hand side of Equation (5), V_j^*, represents the theoretical maximum possible expected payoff from the experiment, if n were to be so large that the Type I and II error rates would be effectively zero ($V^* \equiv \lim_{n\to\infty} V(n) = q\,u(B,H) + (1-q)\,u(A,L)$; Moscarini and Smith 2002). As with setup costs (and for the same reason—lack of data) the value of V_j^* cannot be estimated.

The second term in Equation (5) represents a negative deviation from the maximum expected payoff that approaches zero as n grows larger (hence $\widehat{V}_j(n)$ is a strictly increasing function of n). This term's impact on $\widehat{V}_j(n)$ depends on: 1) the latent parameters ρ and κ, which characterize (in reduced form) the sensitivity of expected payoffs to larger samples relative to the money scale established by $C(n)$; and 2) the variable ϵ_j, an idiosyncratic component of utility that allows the expected payoff from experiment j to differ from the researcher's typical experiment.

Although the parameters ρ and κ lack a structural interpretation and are not the focus of the empirical analysis, a brief description follows. Moscarini and Smith (2002) refer to $1/\rho$ as an index of the experiment's "efficiency," such that, all else equal, experiments with higher values of $1/\rho$ have less need for bigger samples. The quantity κe^{ϵ_j} reflects other factors influencing payoffs, including the utility difference between correct inference and Type I/II errors, the experiment's measurement sensitivity, and the researcher's prior beliefs. The κ term in particular reflects the researcher's global sensitivity to payoffs/costs. Equation (5) is very flexible and can reflect a wide range of behaviors. For example, a high value of κ corresponds with expected payoffs that are flat over a wide range of sample sizes, and thus characterizes a researcher who seems to choose sample sizes at random (i.e., one who is apparently insensitive to costs).

4.3 Heterogeneity

To accommodate heterogeneity in the types of experiments conducted, parameters that vary by study are denoted ρ_j, κ_j, and λ_j (to simplify the discussion, θ refers generically to any parameter

in the set $\{\rho, \kappa, \lambda\}$). Because heterogeneity in experiments is determined in part by heterogeneity in the researchers who design and conduct them, the latter are discussed before presenting the full specification for experimental heterogeneity.

4.3.1 Researchers

Researchers are indexed by i. For each researcher i, there is a "typical" experiment characterized by expected payoffs and costs according to the researcher's values of $\dot{\rho}_i$, $\dot{\kappa}_i$, and $\dot{\lambda}_i$. Recall that researchers are grouped according to their institutional affiliations or student status. Values of $\dot{\theta}_i$'s are likely to vary systematically across these groups, much in the same way researchers from different disciplines vary in the characteristics of their typical experiments. Letting $g(i)$ index the segment for researcher i, the following prior distributions for the $\dot{\theta}_i$'s are defined.

$$\text{logit}^{-1}(\dot{\rho}_i) \,|\bar{\rho}_{g(i)} \sim N\left(\text{logit}^{-1}\left(\bar{\rho}_{g(i)}\right), \tau_\rho^2\right) \tag{6}$$

$$\log(\dot{\kappa}_i) \,|\bar{\kappa}_{g(i)} \sim N\left(\log\left(\bar{\kappa}_{g(i)}\right), \tau_\kappa^2\right) \tag{7}$$

$$\log\left(\dot{\lambda}_i\right) |\bar{\lambda}_{g(i)} \sim N\left(\log\left(\bar{\lambda}_{g(i)}\right), \tau_\lambda^2\right) \tag{8}$$

The prior expected values of researchers' parameters ($\dot{\theta}_i$) are functions of segment-level parameters, denoted $\bar{\theta}_{g(i)}$. These segment-level parameters are then distributed around common parameters, denoted $\bar{\rho}$, $\bar{\kappa}$, and $\bar{\lambda}$, as described in Appendix C.

Recall that more than one researcher can be associated with the same experiment. Given a team of R researchers with indexes represented by the set $\mathcal{R}$, the typical experiment for this team is defined by averaging over the individual collaborators' $\{\dot{\theta}_i\}$'s for $i \in \mathcal{R}$. To allow for flexibility in how this team average reflects heterogeneity among the researchers involved, this average is specified as the generalized mean of the $\{\dot{\theta}_i\}$'s:

$$f_\theta(\mathcal{R}) = \left(\frac{1}{R}\sum_{i\in\mathcal{R}} \dot{\theta}_i^{\gamma_\theta}\right)^{\frac{1}{\gamma_\theta}}, \qquad \gamma_\theta \neq 0, \quad \theta \in \{\rho, \kappa, \lambda\} \tag{9}$$

The value of $f_\theta(\mathcal{R})$ will be closer to the maximum of the $\{\dot{\theta}_i\}$'s if $\gamma_\theta > 1$, closer to the minimum if $\gamma_\theta < 1$, and the simple average if $\gamma_\theta = 1$. Because the $\dot{\theta}_i$'s are unobserved, they are estimated with the other model parameters using a data augmentation approach (Tanner and Wong 1987) and

integrated over during later analysis.

4.3.2 Experiments

Each experiment j has values of ρ_j, κ_j, and λ_j defined as follows.

$$\text{logit}^{-1}(\rho_j) = \text{logit}^{-1}\left(f_\rho(\mathcal{R}_j)\right) + \text{paid}_j\beta_{\rho,\text{paid}} + \text{time}_j\beta_{\rho,\text{time}} \quad (10)$$

$$\log(\kappa_j) = \log\left(f_\kappa(\mathcal{R}_j)\right) + \text{paid}_j\beta_{\kappa,\text{paid}} + \text{time}_j\beta_{\kappa,\text{time}} \quad (11)$$

$$\log(\lambda_j) = \log\left(f_\lambda(\mathcal{R}_j)\right) + \text{paid}_j\beta_{\lambda,\text{paid}} + \text{time}_j\beta_{\lambda,\text{time}} + \text{other}_j\beta_{\lambda,\text{other}} \quad (12)$$

There are many elements common to all three equations. First, the value of θ_j depends in part on the value of $f_\theta(\mathcal{R}_j)$ just discussed. Second, the value of θ_j also depends on two observable characteristics of experiment j. The first is "paid," a dummy variable (coded as $\{-.5, .5\}$) indicating whether experiment j was conducted entirely in the paid pool. The second is "time," the number of (median-centered) hours needed to collect data from one study participant. Finally, λ_j, the cost of time in the lab, also depends on "other," the average number of experiments sharing the lab with study j each day.

4.3.3 Other model specifications

Five versions of the model are estimated, each with varying degrees of observed and unobserved heterogeneity in experiments and researchers:

- **Simple:** The θ_j's (for $\theta \in \{\rho, \kappa, \lambda\}$) are the same for all experiments. Hence $\theta_j = \bar{\theta}$.
- **R:** Researchers have their own $\dot{\theta}_i$'s, but these are determined exactly by their segment membership—i.e., $\dot{\theta}_i = \theta_{g(i)}$. The $\dot{\theta}_i$'s for teams of collaborating researchers mix according to the $f_\theta(\mathcal{R}_j)$ functions defined by Equation (9), but there are no observed experimental characteristics included in the θ_j's. Hence $g(\theta_j) = g\left(f_\theta(\mathcal{R}_j)\right)$ where $g(\cdot)$ represents the appropriate log or inverse-logit transformation.
- **E:** The θ_j's do not include the researcher effects, $f_\theta(\mathcal{R}_j)$. Hence $g(\theta_j) = x_j\beta_\theta$ where x_j

indicates the observed experimental variables "paid," "time," and "other" in Equations (10)–(12).

- **R+E:** Both features specified in Models R and E are included, and Equations (10)–(12) are unmodified. Researchers' $\dot{\theta}_i$'s, however, do not differ within segments (i.e., $\dot{\theta}_i = \theta_{g(i)}$).
- **Full:** This is the specification presented in the main text (i.e., the $\dot{\theta}_i$'s are distributed normally around their segment means with prior variances τ_θ^2).

The five models are summarized in Table 2.

4.4 Estimation

Appendix C presents the Bayesian hyper-prior and posterior distributions of the model parameters, and the counterfactual procedure. Here a brief sketch of the model's likelihood function (i.e., the likelihood of the parameters, conditional on the observed sample sizes n) is given.

The likelihood function is predicated on the optimality of the observed sample sizes given the researchers's expected payoffs and costs. Conditional on the model parameters, and given that the sample size for experiment j was chosen to be n_j (and not $n_j + 1$ or $n_j - 1$), there is a limited range of ϵ_j's that can rationalize n_j as having provided the greatest net expected value to the researcher(s). The likelihood of n_j is accordingly defined as the total probability of ϵ_j within this valid range (see, e.g., Lee and Allenby 2014). The distribution of the ϵ_j's is assumed to be:

$$\epsilon_j \sim N\left(0, \sigma^2\right) \tag{13}$$

Note that although n_j is a discrete variable, the resulting likelihood function is continuous in the model parameters, conditional on the n_j's.

5 Parameter Estimates

Five versions of the model are estimated, each allowing for different degrees of observed and unobserved heterogeneity in experiments and researchers. Table 2 indicates the RMSE of posterior

Table 2 Model Specifications and Fit

	MODEL				
	Simple	*R*	*E*	*R+E*	*Full*
Source of heterogeneity					
Observed researcher characteristics ($\bar{\theta}_g$, γ_θ)		x		x	x
Observed experiment characteristics (β_θ)			x	x	x
Unobserved researcher characteristics (τ^2_θ)					x
RMSE of posterior predictions (%)	78.0	77.2	75.6	74.8	64.6

Table 3 Parameter Summary

Type	*Parameter*	*Description*
Cost	α	Translates credit reimbursements to money scale
	δ	Incremental cost of each day in lab
	$\bar{\lambda}$	Prior mean of researchers' log lab cost
	τ^2_λ	Prior variance of researchers' log lab cost
	$\beta_{\lambda,\text{time}}$	Difference in log lab cost from 1 hour in duration
	$\beta_{\lambda,\text{paid}}$	Difference in log lab cost from using only paid pool
	$\beta_{\lambda,\text{other}}$	Difference in log lab cost from 1 additional study
Payoff	σ	Scale of idiosyncratic payoff (ϵ_j)
	$\bar{\kappa}$, $\bar{\rho}$	Prior mean of researchers' payoff sensitivity
	τ^2_κ, τ^2_ρ	Prior variance of researchers' payoff sensitivity
	$\beta_{\kappa,\text{time}}$, $\beta_{\rho,\text{time}}$	Difference in payoff sensitivity from 1 hour in duration
	$\beta_{\kappa,\text{paid}}$, $\beta_{\rho,\text{paid}}$	Difference in payoff sensitivity from using only paid pool
Mixing	γ_λ	Day cost mixing parameter
	γ_κ, γ_ρ	Payoff sensitivity mixing parameters

predictions for each model. Because the full model specification presented in the previous section fits the observed sample sizes best, and because the choice of model specification does not have a significant qualitative impact on counterfactual results (see Appendix E), all results discussed here pertain to the full model. Posterior estimates of parameters for all specifications are shown in Table 4.

5.1 Cost Parameters

Results related to the two types of cost (reimbursing participants and spending time in the lab) are presented first.

Table 4 Parameter Estimates

		MODEL				
Type	*Parameter*	*Simple*	*R*	*E*	*R+E*	*Full*
Cost	α	2.72	4.55	6.44	7.67	6.66
		(2.03, 3.57)	(3.26, 6.19)	(3.80, 9.29)	(5.17, 11.36)	(4.52, 9.92)
	δ	0.29	0.29	0.29	0.38	0.28
		(0.18, 0.41)	(0.16, 0.42)	(0.17, 0.43)	(0.22, 0.58)	(0.17, 0.43)
	$\log \bar{\lambda}$	−0.81	−0.38	−0.30	−1.53	0.14
		(−1.40, −0.33)	(−1.42, 0.74)	(−1.11, 0.46)	(−2.87, −0.03)	(−0.82, 1.08)
	τ_λ^2					6.36
						(3.84, 9.53)
	$\beta_{\lambda,\text{time}}$			2.19	2.56	2.45
				(1.69, 2.68)	(1.74, 3.42)	(1.50, 3.37)
	$\beta_{\lambda,\text{paid}}$			0.20	1.81	−0.40
				(−0.31, 0.79)	(1.14, 2.40)	(−1.07, 0.23)
	$\beta_{\lambda,\text{other}}$			0.13	0.24	−0.02
				(0.03, 0.24)	(0.10, 0.43)	(−0.16, 0.14)
Payoff	σ	1.80	1.77	1.64	1.65	1.40
		(1.65, 1.98)	(1.61, 1.98)	(1.52, 1.79)	(1.50, 1.83)	(1.28, 1.55)
	$\log \bar{\kappa}$	8.67	6.97	8.74	6.91	6.88
		(8.43, 8.91)	(6.19, 7.87)	(8.41, 9.00)	(6.20, 7.64)	(5.88, 7.75)
	$\text{logit}^{-1}(\bar{\rho})$	4.47	3.25	4.43	3.33	3.45
		(4.29, 4.66)	(2.40, 4.17)	(4.24, 4.68)	(2.61, 4.24)	(2.54, 4.35)
	τ_κ^2					0.70
						(0.43, 1.07)
	τ_ρ^2					0.36
						(0.21, 0.68)
	$\beta_{\kappa,\text{time}}$			0.84	1.04	0.87
				(0.36, 1.25)	(0.59, 1.55)	(0.36, 1.44)
	$\beta_{\rho,\text{time}}$			0.94	0.46	0.51
				(0.31, 1.62)	(−0.08, 0.98)	(−0.29, 1.46)
	$\beta_{\kappa,\text{paid}}$			−1.40	−0.92	−0.91
				(−1.83, −0.86)	(−1.39, −0.43)	(−1.38, −0.41)
	$\beta_{\rho,\text{paid}}$			−0.23	−0.45	−0.41
				(−0.52, 0.11)	(−0.72, −0.12)	(−0.76, 0.00)
Mixing	γ_λ		0.59		−1.08	0.60
			(−1.50, 2.69)		(−2.12, −0.31)	(0.18, 1.83)
	γ_κ		1.21		1.11	2.06
			(−0.69, 3.35)		(−0.59, 3.29)	(0.81, 3.60)
	γ_ρ		1.01		0.73	1.17
			(−0.81, 2.99)		(−1.11, 2.28)	(−0.55, 3.05)

Note. Estimates are posterior means, with Bayesian 95% CI's in parentheses.

5.1.1 Participant reimbursement

The parameter α, which translates course credit to the same scale as cash payments, implies an exchange rate of €6.66 per unit of course credit. Because a 30 minute study is typically reimbursed at either 1 credit or €5, this exchange rate suggests researchers might receive slightly more disutility when paying with credit than when paying with cash (although the 95% CI for α includes €5, so there may in fact be no difference).

Demand for observations in the two subject pools is about equally sensitive to differences in the reimbursement rates. To make this comparison, each experiment's arc elasticity of demand (for participants) is calculated using cash and credit payments that are 5% above and below those in the data (hence the arc elasticities are $E_j = \frac{\Delta n_j}{n_j} \Big/ \frac{\Delta p_j}{p_j}$, with $\Delta p_j = .1 p_j$ and $\Delta n_j = \mathbb{E}\left[n_j | 1.05 p_j\right] - \mathbb{E}\left[n_j | .95 p_j\right]$). For experiments reimbursing participants exclusively with credits, money, or both, median payment elasticities are $-.50$, $-.51$, and $-.45$, respectively, suggesting if participant compensation was 10% higher, we would have expected the median experiment to use about 5% fewer participants. Consistent with the counterfactual analysis presented later, this result suggests researchers are sufficiently cost sensitive to allow interventions targeting participant reimbursements to have meaningful effects on sample sizes.

5.1.2 Days in the lab

The estimate for δ is less than 1, indicating each additional day in the lab is less costly than the previous. Disutility from time in the lab depends on both δ and λ_j, and estimates for these parameters indicate that adding an extra day to the median experiment would generate the same disutility as paying an extra €1.28 to one of the participants—that is, almost no disutility at all. Because researchers schedule lab time in advance of the study, this estimate reflects researchers' anticipated disutility from time in the lab, and not its realization, which might explain why this median is so low. Nevertheless, even though the median is low, this cost is highly variable across experiments. In about 3% of experiments, collecting the $n + 1^{\text{th}}$ observation would have required another day in the lab *and* that extra day would have generated disutility exceeding $5\, p_j$. Hence there is a small

subset of experiments for which time in the lab might have provided a meaningful disincentive against working with bigger samples.

Environmental factors and experimental characteristics also bear on the cost of time spent in the lab. First, $\beta_{\lambda,\text{other}}$ is very close to 0, suggesting that on average, time costs are not meaningfully related to the number of experiments sharing the lab each day. Second, $\beta_{\lambda,\text{time}}$ is positive, and for each 15 minute difference in duration, average daily lab costs are about 33% higher (but against a low baseline, as noted above). Third, experiments that paid course credit have the highest overall costs, both in terms of lab time and participant reimbursement.

5.2 Payoff Parameters

The parameters concerning payoffs lack structural interpretation, nevertheless two patterns emerge from their estimates: 1) experiments taking longer to administer have payoff characteristics ($\beta_{\kappa,\text{time}}$) related to bigger samples; and 2) experiments run exclusively in the paid pool have payoff characteristics ($\beta_{\kappa,\text{paid}}$ and $\beta_{\rho,\text{paid}}$) related to smaller samples. Unfortunately it is not possible to pinpoint the exact reasons for these results with the available data.

5.3 Mixing Parameters for Teams of Researchers

Recall the γ parameters describe how collaborating researchers' values of $\dot{\lambda}_i$, $\dot{\kappa}_i$, and $\dot{\rho}_i$ mix to characterize a team of researchers' typical experiment. The estimates for γ_λ and γ_κ are consistent with collaborators pooling their resources to reduce costs. First, time costs are weighted in favor of the team member with the lowest value of $\dot{\lambda}_i$. Although the data do not say which researcher(s) were present in the lab when collecting data, the low estimate for γ_λ, coupled with the fact that the student researcher segment has the lowest estimated value of $\bar{\lambda}_g$, suggests the involvement of students in lab-based research may be beneficial for sample sizes. Second, net expected payoffs are weighted in favor of the team member with the greatest incentive for bigger samples ($\dot{\kappa}_i$), which is consistent with unrelated experiments sharing a sample to decrease costs (in such cases we would expect the combined payoffs to reflect those of the experiment with the greatest sensitivity to bigger

samples).

6 Manipulating Costs to Incentivize Bigger Samples

The empirical results suggest institutional policies that lower researchers' costs might have a meaningful impact on sample sizes. However, it is important not only to identify which interventions will be most effective, but also to predict their likely effects *before* implementation. It would be ideal if the institutions responsible for funding and operating research facilities could conduct pilot studies to identify which policies to implement on a wider scale. Unfortunately, in the real world such experimentation—even on a small scale—can have serious, unintended, and negative consequences for the institution and individual researchers involved (Ioannidis 2012a). Counterfactual analysis, however, provides a useful and feasible alternative.

Hence, with the goal of assessing the impact of potential cost-lowering interventions prior to implementation, the estimates from the full model are used to simulate the effects of these interventions on sample sizes at this lab (and more specifically, to estimate the magnitude of any improvements). Furthermore, by joining these results with survey data describing typical effect and sample sizes, the expected magnitude of improvements in experimental power due to these bigger samples can be estimated as well. The cost interventions are presented first, followed by the simulation results.

6.1 Counterfactual Policies

Two policies designed to lower the disutility from reimbursing participants are presented here, and the general simulation procedure is described in Appendix C.[3] Under both reimbursement policies, the amount received by each participant would not change. However, by manipulating the impact of remuneration on researchers' budgets, the disutility from compensating participants would be diminished, reducing the incentive against collecting a bigger sample. These simulations measure

[3] A third simulation based on expanding the lab's capacity is reported in Appendix D. Consistent with the high capacity of the lab and low estimated costs of spending time there, this simulation points to negligible gains from implementing this policy.

the impact of these policies on the set of studies observed in the data, and hence do not speak to how the policies might have led to greater or fewer numbers of experimented conducted. Moreover, it is assumed for the sake of these simulations that these policies would not have affected experimental designs. For example, a study depending on a between-subjects manipulation cannot change to a within-subjects manipulation in the context of these counterfactual simulations. Further details about each counterfactual simulation are described next.

6.1.1 Credits intervention

The "credits" intervention targets experiments in the credit pool by raising each student's requirement for and each researcher's allotment of lab-based course credit by $r\%$. This policy is simulated by changing Equation (4) to

$$p_j(s) = \text{euros}_j + (1 - s)\,\alpha\text{credits}_j, \tag{14}$$

where $s = r/(1 + r)$ is the size of the subsidy generated by this policy. Hence increasing each researchers' credit budget (and each student's credit requirement) by $r = 50\%$ would decrease the disutility of reimbursing participants with credit by $s = 1/3$. This simulation assumes that students who previously participated two 30 minute studies would be willing to participate in three under the new policy (with $r = 50\%$), and that some of the students who currently do not obtain credit would seek it out under the new, higher requirements.

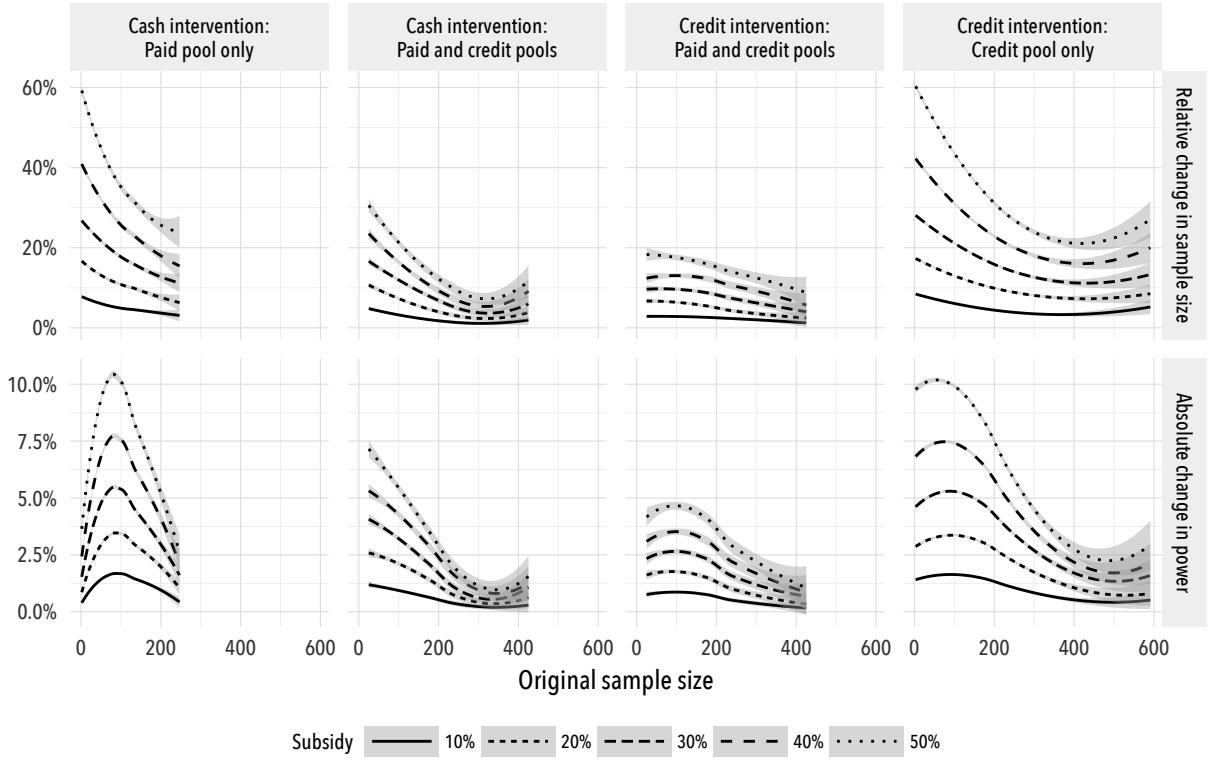
6.1.2 Cash intervention

The "cash" intervention mirrors the first, but targets the paid pool. Study participants are compensated at exactly the same rate, but the lab directly subsidizes this payment at a rate of $s\%$.

$$p_j(s) = (1 - s)\,\text{euros}_j + \alpha\,\text{credits}_j \tag{15}$$

Hence, a researcher paying a participant €5 would incur disutility as if the payment were only $(1 - s) \times$€5. This simulation assumes the standard rate of €10/hour does not change, and that there are enough participants to satisfy any higher demand.

Figure 6 Changes in Sample Size (top row) and Experimental Power (bottom row) under Counterfactual Reimbursement Policies



Notes. The top row shows the expected percent increase in sample size, and the bottom row the expected difference in experimental power. The left columns depict changes under the "cash" and the right columns changes under the "credit" interventions. Lines indicate estimates from a LOESS regression of simulated outcomes on sample size, with 95% confidence bounds in shown grey.

6.2 Results

The simulation results (based on parameter estimates from the full model) show that both reimbursement policies can reduce the incentive against collecting bigger samples (results are qualitatively similar for other model specifications; details are provided in Appendix E). The expected impacts of these interventions on both sample sizes and experimental power are considered first, followed by an analysis of the total costs of implementing these policies.

6.2.1 Implications for sample sizes

Figure 6 shows the main results of the cash and credit interventions for various levels of the subsidy s. As expected, bigger subsidies lead to bigger increases in sample size, with the largest effects

among experiments conducted exclusively in the targeted pool (the outside columns in Figure 6).

The magnitude of the increases in sample size (the top row in Figure 6) varies across experiments, with those originally run with the smallest samples generating the greatest (relative) increases. But even among mid-sized experiments, both subsidies produce substantial improvements. Specifically, among the 82 (251) studies run exclusively in the paid (credit) pool with original sample sizes between 50 and 200, a 50% subsidy leads to an average sample size that is expected to be about 34% (43%) higher. Although expected increases among studies using both participant pools are smaller, they are still substantial in absolute terms: A 50% subsidy of either credit or cash (but not both) yields an expected sample size increase of about 20%. All of these expected increases in sample size lead to decreases in total outlay from the researchers' accounts, which has implications for the total cost of the policies as discussed below.

6.2.2 Implications for experimental power

Although measuring the impact of cost interventions on sample sizes is the main objective of this paper, the goal of incentivizing bigger samples is motivated by the belief that experimental power is generally too low. It would thus be useful to know the extent to which the expected increases in sample size translate into lower Type II error rates. Unfortunately, the lab archival data contain insufficient information about experiments to infer expected Type II rates.

A proxy for these data is available, however, in the form of survey responses collected by Gervais et al. (2015), which are used to estimate the average power for typical studies of different sample sizes. This then permits a rough estimate of the expected change in experimental power under the counterfactual policies.

The survey data comprise 135 responses from members of the Society of Experimental Social Psychology (population: 937) who answered questions about their typical experiments (much of the research conducted at this lab follows paradigms closely related to those found in social psychology). More specifically, the survey responses include two variables of interest: 1) the typical number of participants in a two-cell experiment, and 2) the typical effect size for such an

experiment. Gervais et al. (2015)'s procedure is used to calculate the expected power for these hypothetical experiments. These values are then regressed on the log of the stated sample sizes (McFadden's pseudo-R^2 for logistic regression: .37), leading to estimated Type II error rates for typical experiments with n observations.

The bottom row of Figure 6 shows the absolute difference in expected experimental power corresponding with the sample size increases depicted in the top row. It is difficult to assess the accuracy of the estimated increases in power, as they are based on data describing different groups of researchers. Nevertheless, the estimates are not entirely without value, and point to two interesting patterns.

First, the large expected increases in sample size among smaller experiments do not necessarily translate to large expected increases in experimental power. This is likely due to small effect sizes in the social sciences combined with the shape of t-test's power curve. The second conclusion is that the largest increases in power occur within the range of the most typical sample sizes (i.e., those originally run with roughly 50–200 participants). Notably, calls for larger samples in the experimental social sciences often explicitly mention 2- and 4-cell studies with fewer than 50 participants as being too small (e.g., Simmons 2014). Hence, this intervention is expected to achieve the largest gains among the subset of studies that might benefit the most. Within this group of studies, expected power is estimated to be .10 higher for studies conducted exclusively in one pool (and about half that among studies using both pools), compared to .08 higher among studies with less than 50 participants.

6.2.3 Cost of implementation

Of the two policies considered here, the credit intervention would be the least expensive to implement, as course credit is free to produce. Although the money cost of expanding credit requirements is nil, there may be institutional barriers standing in the way of this policy (previous efforts to expand the credit pool at this lab were unsuccessful).

The cash intervention appears to be highly efficient in spite of its non-zero implementation

cost. Among studies using the paid pool (either partially or exclusively), a 50% subsidy would be expected to increase sample sizes by about 36%. To obtain this increase in sample size, total reimbursements to participants (from both the lab and individual researchers' accounts) would have needed to grow by about 32%, or €610 per month (when averaged over the 63 months in the data).

Under this 50% cash subsidy, researchers would be expected to commit about 32% less of their available research budgets to the set of experiments described in the archive. And some, or perhaps most of this savings would presumably be reinvested in running additional experiments. Because limitations in the archive data preclude modeling researchers' decisions to run experiments and individuals' decisions to participate, it is difficult to say how many more experiments would have been conducted under the counterfactual policy. The estimate of €610 should therefore be considered a lower bound on the true monthly cost to the institution.

An implication of this analysis is that funding institutions, which often have stated or implicit goals of improving the way their researchers carry out their work, can modify researchers' incentives in ways that can bring about such improvements. However, implementing such policies may require the commitment of additional financial resources on the part of the funding institution. That is, institutions wanting better scientific outcomes should recognize they might need to equip their researchers with additional resources.

7 Conclusion

Limitations on researcher's resources can lead to them making choices that are poorly aligned with the goals of their stakeholders, or even in conflict their own ideals of how other researchers ought to carry out their work. We should therefore seek to understand how researchers' incentives affect their choices, as this knowledge will allow us to design interventions that can elicit more normatively desirable behaviors.

This paper seeks to further our understanding of researchers' incentives by providing the first empirical measure of how researchers' costs affect their chosen sample sizes, and by simulating how cost-reducing policies might have affected them. The method used for this analysis is flexible

and general, and as such can serve as a template for similar analyses at other institutions (including online subject pools, e.g. Amazon MTurk).

Across a wide range of experiments and researchers, the need to reimburse participants generates a significant disincentive against working with bigger samples. At the same time, the strength of this disincentive suggests manipulating researchers' costs could bring about meaningful improvements in sample size and experimental power. Indeed, whether researchers pay participants in course credit or cash, a policy of (further) subsidizing participant reimbursement might be an efficient strategy for increasing sample sizes. Given the long standing problem of low-powered experiments in the social sciences, as well as more recent attention on scientific replication and reliability, such improvements might have a positive impact on scientific outcomes and lead to more efficient use of research funds.

An important caveat to the empirical results presented earlier is that they pertain only to the Erasmus Behavioral Lab and the researchers who rely on it. Parameter estimates will differ across institutions due to differences across populations of researchers and the resources available to them. As such, these results should not be expected to generalize immediately to other populations of experimental social scientists. But by combining data from multiple labs, future research might consider how differences in research environments correlate with differences in cost sensitivity and scientific outcomes.

In spite of this limitation, the results presented here show the potential for cost-reducing policies to bring researchers' incentives into closer alignment with those of their stakeholders, and generate meaningful improvements in scientific outcomes. From a methodological perspective, this study demonstrates the value of combining empirical quantitative methods from consumer research with archival data in order to gain a better understanding of how researchers conduct their work. Future work in this area will hopefully consider other settings and empirical methods (e.g., field studies).

Finally, although the goal of increasing sample sizes enjoys broad support within the scientific community, there are other ways to increase experimental power (Maxwell 2004; Abraham and Russell 2008) not considered here. Moreover, even though improvements in experimental power

are expected to contribute to better scientific outcomes, these improvements by themselves cannot solve all of the problems of reliability and replicability currently facing the experimental social sciences. This study however demonstrates the value of considering researchers' decisions in a consumption framework as a way of identifying new solutions to old problems.

References

Abraham, W. T., and D. W. Russell. 2008. "Statistical power analysis in psychological research." *Social and Personality Psychology Compass* 2 (1): 283–301. doi:10.1111/j.1751-9004.2007.00052.x.

Allison, D. B., R. L. Allison, M. S. Faith, F. Paultre, and F. X. Pi-Sunyer. 1997. "Power and money: Designing statistically powerful studies while minimizing financial costs." *Psychological Methods* 2 (1): 20–33. doi:10.1037/1082-989X.2.1.20.

Asendorpf, J. B., M. Conner, F. De Fruyt, J. De Houwer, J. J. Denissen, K. Fiedler, S. Fiedler, et al. 2013. "Recommendations for increasing replicability in psychology." *European Journal of Personality* 27 (2): 108–119. doi:10.1002/per.1919.

Bakker, M., A. van Dijk, and J. M. Wicherts. 2012. "The rules of the game called psychological science." *Perspectives on Psychological Science* 7 (6): 543–554. doi:10.1177/1745691612459060.

Baumeister, R. F. 2016. "Charting the future of social psychology on stormy seas: Winners, losers, and recommendations." *Journal of Experimental Social Psychology* In press. doi:10.1016/j.jesp.2016.02.003.

Blattberg, R. C. 1979. "The design of advertising experiments using statistical decision theory." *Journal of Marketing Research* 16 (2): 191–202. doi:10.2307/3150683.

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. "Power failure: Why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–376. doi:10.1038/nrn3475.

Chatterjee, R., J. Eliashberg, H. Gatignon, and L. M. Lodish. 1988. "A practical Bayesian approach to selection of optimal market testing strategies." *Journal of Marketing Research:* 363–375. doi:10.2307/3172947.

Cohen, J. 1962. "The statistical power of abnormal-social psychological research: A review." *Journal of Abnormal and Social Psychology* 65 (3): 145–153. doi:10.1037/h0045186.

Cohen, J. 1969. *Statistical power analysis for the bebehavior sciences.* 1st. San Diego: Academic Press.

Cohen, J. 1992a. "A power primer." *Psychological Bulletin* 112 (1): 155–159. doi:10.1037/0033-2909.112.1.155.

Cohen, J. 1992b. "Statistical power analysis." *Current Directions in Psychological Science* 1 (3): 98–101. doi:10.1111/1467-8721.ep10768783.

Dasgupta, P., and P. A. David. 1994. "Toward a new economics of science." *Research Policy* 23 (5): 487–521. doi:10.1016/0048-7333(94)01002-1.

Fiedler, K., F. Kutzner, and J. I. Krueger. 2012. "The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate." *Perspectives on Psychological Science* 7 (6): 661–669. doi:10.1177/1745691612462587.

Gelman, A., and J. Carlin. 2014. "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors." *Perspectives on Psychological Science* 9 (6): 641–651. doi:10.1177/1745691614551642.

Gervais, W. M., J. A. Jewell, M. B. Najle, and B. K. Ng. 2015. "A powerful nudge? Presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs." *Social Psychological and Personality Science* 6 (7): 847–853. doi:10.1177/1948550615584199.

Ginter, J. L., M. C. Cooper, C. Obermiller, and T. J. Page Jr. 1981. "The design of advertising experiments using statistical decision theory: An extension." *Journal of Marketing Research* 18 (1): 120–123. doi:10.2307/3151323.

Gneezy, U., A. Imas, and K. Madarász. 2014. "Conscience accounting: Emotional dynamics and social behavior." *Management Science* 60 (11): 2645–2658. doi:10.1287/mnsc.2014.1942.

Ioannidis, J. P. 2005. "Why most published research findings are false." *PLoS Medicine* 2 (8): 696–701. doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. 2012a. "Scientific communication is down at the moment, please check again later." *Psychological Inquiry* 23 (3): 267–270. doi:10.1080/1047840X.2012.699427.

Ioannidis, J. P. 2012b. "Why science is not necessarily self-correcting." *Perspectives on Psychological Science* 7 (6): 645–654. doi:10.1177/1745691612464056.

Lee, S., and G. M. Allenby. 2014. "Modeling indivisible demand." *Marketing Science* 33 (3): 364–381. doi:10.1287/mksc.2013.0829.

Lenth, R. V. 2001. "Some practical guidelines for effective sample size determination." *The American Statistician* 55 (3): 187–193. doi:10.1198/000313001317098149.

Marszalek, J. M., C. Barber, J. Kohlhart, and C. B. Holmes. 2011. "Sample size in psychological research over the past 30 years." *Perceptual and motor skills* 112 (2): 331–348. doi:10.2466/03.11.PMS.112.2.331-348.

Maxwell, S. E. 2004. "The persistence of underpowered studies in psychological research: Causes, consequences, and remedies." *Psychological Methods* 9 (2): 147–163. doi:10.1037/1082-989X.9.2.147.

Maxwell, S. E., K. Kelley, and J. R. Rausch. 2008. "Sample size planning for statistical power and accuracy in parameter estimation." *Annual Review of Psychology* 59:537–563. doi:10.1146/annurev.psych.59.103006.093735.

McClelland, G. H. 2000. "Increasing statistical power without increasing sample size." *American Psychologist* 55 (8): 963–964. doi:10.1037/0003-066X.55.8.963.

Meyer, R. J. 2015. "Editorial: A field guide to publishing in an era of doubt." *Journal of Marketing Research* 52 (5): 577–579. doi:10.1509/jmr.52.5.577.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting transparency in social science research." *Science* 343 (6166): 30–31. doi:10.1126/science.1245317.

Moscarini, G., and L. Smith. 2002. "The law of large demand for information." *Econometrica* 70 (6): 2351–2366. doi:10.1111/j.1468-0262.2002.00442.x.

Nosek, B. A., J. R. Spies, and M. Motyl. 2012. "Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability." *Perspectives on Psychological Science* 7 (6): 615–631. doi:10.1177/1745691612459058.

Pearl, J. 2009. *Causality: Models, reasoning, and inference*. 2nd. Cambridge University Press.

Rosenthal, R. 1979. "The file drawer problem and tolerance for null results." *Psychological Bulletin* 86 (3): 638–641. doi:10.1037/0033-2909.86.3.638.

Sawyer, A. G., and A. D. Ball. 1981. "Statistical power and effect size in marketing research." *Journal of Marketing Research* 18 (3): 275–290. doi:10.2307/3150969.

Schimmack, U. 2012. "The ironic effect of significant results on the credibility of multiple-study articles." *Psychological Methods* 17 (4): 551. doi:10.1037/a0029487.

Scutari, M. 2010. "Learning Bayesian networks with the bnlearn R package." *Journal of Statistical Software* 35 (3): 1–22. doi:10.18637/jss.v035.i03.

Sedlmeier, P., and G. Gigerenzer. 1989. "Do studies of statistical power have an effect on the power of studies?" *Psychological Bulletin* 105 (2): 309–316. doi:10.1037/0033-2909.105.2.309.

Shen, W., T. B. Kiger, S. E. Davies, R. L. Rasch, K. M. Simon, and D. S. Ones. 2011. "Samples in applied psychology: Over a decade of research in review." *Journal of Applied Psychology* 96 (5): 1055. doi:10.1037/a0023322.

Simmons, J. 2014. "MTurk vs. The Lab: Either Way We Need Big Samples." Data Colada. April 4. http://web.archive.org/web/20170313102117/http://datacolada.org/18.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22 (11): 1359–1366. doi:10.1177/0956797611417632.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2013. "Life after ***P***-hacking." Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. doi:10.2139/ssrn.2205186.

Simonsohn, U., L. D. Nelson, and J. P. Simmons. 2014. "***P***-curve: A key to the file-drawer." *Journal of Experimental Psychology: General* 143 (2): 534–547. doi:10.1037/a0033242.

Stan Development Team. 2015. *CmdStan: The command-line interface to Stan, Version 2.8*. http://mc-stan.org/cmdstan.html.

Stephan, P. E. 1996. "The economics of science." *Journal of Economic Literature* 34 (3): 1199–1235. https://www.jstor.org/stable/2729500.

Tanner, M. A., and W. H. Wong. 1987. "The calculation of posterior distributions by data augmentation." *Journal of the American Statistical Association* 82 (398): 528–540. doi:10.1080/01621459.1987.10478458.

VanVoorhis, C. R. W., and B. L. Morgan. 2007. "Understanding power and rules of thumb for determining sample sizes." *Tutorials in Quantitative Methods for Psychology* 3 (2): 43–50. doi:10.20982/tqmp.03.2.p043.

Winkens, B., H. J. Schouten, G. J. van Breukelen, and M. P. Berger. 2006. "Optimal number of repeated measures and group sizes in clinical trials with linearly divergent treatment effects." *Contemporary Clinical Trials* 27 (1): 57–69. doi:10.1016/j.cct.2005.09.005.

A Sample Selection Procedure

After identifying and linking records for those using both participant pools, the archive contains 782 experimental records. The following steps lead to the final estimation sample of 683.

1. Each session in a "multipart" experiment (i.e., one that takes repeated measures of the same sample over time) is registered as a separate experiment in the archive. Only the first session is used for estimation, eliminating 21 observations.

2. Experiments that were not conducted in the lab facilities, such as those marked as taking place online, at the nearby medical center, or at a local movie theater, are excluded, eliminating another 25 observations.

3. Only experiments for which 2 or more participants registered can be used for estimation (see Appendix C), eliminating another 53 observations. The data contain a continuum of sample sizes between 1 and 591, with no clear line separating pre-tests and aborted studies from "real" experiments. Hence all experiments with $n > 1$ are included.

For each experiment, the following statistics are calculated. First, the time of the last observation collected on the final day of the experiment, the duration of the experiment, and an assumed cut-off time of 5pm jointly determine whether adding another ex ante observation to the experiment would require another day in the lab. Second, for each experiment j that used both pools, the expected payment to the $n+1^{\text{th}}$ participant is defined as $p_j = \omega_j \,\text{euros}_j + (1-\omega_j)\,\alpha\,\text{credits}_j$, where ω_j indicates the share of the first n_j observations collected from the paid pool. Third, for each experiment, the maximum number of observations that could have been collected in a single day is estimated as the larger of two quantities: 1) the most observations collected on any of the first

$D_j(n_j) - 1$ days (or zero if $D_j(n_j) = 1$); or 2) the number of observations collected on the final day divided by the share the last day used (assuming a 9.25 hour day). For example, an experiment run on one day with $n = 10$ and a final observation collected halfway through the day has an estimated maximum number of observations per day of 20.

B Tests for Causal Influence of p_j on n_j

The counterfactual analyses are predicated on the assumption that participant reimbursement levels, p_j, causally influence chosen sample sizes, n_j. This assumption leads to testable predictions about patterns of conditional independence among the estimation data, which, if detected, can positively establish such a causal relationship (Pearl 2009). These tests are carried out using the **ci.test** function in the **bnlearn** R package (Scutari 2010), specifying the **smc-mi-g** mutual information test statistic.

The tests support the assumption that p_j causally affects n_j and lend credence to the validity of the counterfactual analysis. First, conditional on the number of days on which data were collected (D_j), the time needed to collect each observation (time_j) is significantly related to observed sample sizes (i.e., they are conditionally dependent): $\text{time}_j \not\perp\!\!\!\perp n_j \mid D_j$ ($p = .0002$). Second, when also conditioning on p_j, the dependency between time_j and n_j is broken, thus identifying p_j as a mediator between time_j and n_j: $\text{time}_j \perp\!\!\!\perp n_j \mid D_j, p_j$ ($p = .97$). Third, even though this pattern of results is consistent with two causal paths: $\text{time}_j \rightarrow p_j \rightarrow n_j$, and the reverse, $n_j \rightarrow p_j \rightarrow \text{time}_j$, the latter case is highly unlikely because it would require researchers to design their experiments after choosing their sample sizes (Pearl 2009, Definition 2.7.4).

C Estimation Details

The prior and posterior probability distributions of the model parameters are first presented, followed by a discussion of the general approach to the counterfactual simulations.

C.1 Bayesian Prior Distribution

The joint prior distribution of the model parameters is denoted $\pi(\Theta)$, and is defined as the product of the following marginal distributions.

$$\begin{aligned}
\alpha &\sim Ga(5,1) \quad & \delta &\sim Ga(5,5) \quad & \sigma &\sim Inv\text{-}Ga(4,3) \\
\beta_\theta &\sim N(0,1) \quad & \gamma_\theta &\sim N(1,1) \quad & \tau_\theta^2 &\sim Inv\text{-}Ga(4,3) \\
\text{logit}^{-1}\,\bar{\rho}_g|\bar{\rho} &\sim N\left(\text{logit}^{-1}\,\bar{\rho}, 1\right) & & & \text{logit}^{-1}\,\bar{\rho} &\sim N(0,1) \\
\log\bar{\kappa}_g|\bar{\kappa} &\sim N(\log\bar{\kappa}, 1) & & & \log\bar{\kappa} &\sim N(0,1) \\
\log\bar{\lambda}_g|\bar{\lambda} &\sim N\left(\log\bar{\lambda}, 1\right) & & & \log\bar{\lambda} &\sim N(0,1)
\end{aligned} \tag{C.1}$$

C.2 Bayesian Posterior Distribution

The likelihood of the model parameters, conditional on the data, is obtained by assuming that the observed sample size for experiment j had a higher net expected utility than any other alternative value of n_j. Formally, this assumption entails two inequalities:

$$\widehat{V}_j(n_j) - C_j(n_j) \geq \widehat{V}_j(n_j + \nu_j) - C_j(n_j + \nu_j), \qquad \nu_j \in \{1, 2, \dots\} \tag{C.2}$$

$$\widehat{V}_j(n_j) - C_j(n_j) \geq \widehat{V}_j(n_j - \mu_j) - C_j(n_j - \mu_j), \qquad \mu_j \in \{1, 2, \dots, n_j - 1\} \tag{C.3}$$

The first (second) inequality ensures the researcher could not have increased the expected net payoff from the experiment by an ex ante increase (decrease) in sample size. The institutional guarantee on participant reimbursement ensures these values of n_j were feasible alternatives.

To derive the likelihood function, substitute Equations (3–5) into (C.2) and (C.3) and rearrange terms to isolate ϵ_j, which leads to the following two inequality conditions for ϵ_j (see, e.g., Lee and Allenby 2014):

$$b_j^{\ell}(n_j) \leq \epsilon_j \qquad \text{and} \qquad \epsilon_j \leq b_j^{u}(n_j) \tag{C.4}$$

These upper and lower bounds are defined for positive integers ν_j and μ_j as:

$$b_j^u(n_j) \equiv \min_{\nu_j \geq 1} \left(-\log \kappa_j - \log \left\{ \frac{\rho_j^{n_j}}{\sqrt{n_j}} - \frac{\rho_j^{n_j+\nu_j}}{\sqrt{n_j+\nu_j}} \right\} + \log \left\{ \lambda_j \left[D_j(n_j+\nu_j)^{\delta} - D_j(n_j)^{\delta} \right] + \nu_j p_j \right\} \right) \tag{C.5}$$

$$b_j^{\ell}(n_j) \equiv \max_{n_j > \mu_j \geq 1} \left(-\log \kappa_j - \log \left\{ \frac{\rho_j^{n_j-\mu_j}}{\sqrt{n_j-\mu_j}} - \frac{\rho_j^{n_j}}{\sqrt{n_j}} \right\} + \log \left\{ \lambda_j \left[D_j(n_j)^{\delta} - D_j(n_j-\mu_j)^{\delta} \right] + \mu_j p_j \right\} \right) \tag{C.6}$$

The values of ν_j and μ_j that minimize/maximize the upper and lower bounds are found via numerical search. Note that V_j^* and F_j appear on both the right and left-hand sides of (C.2) and (C.3) and thus cancel out of Equations (C.5) and (C.6). These quantities therefore cannot be estimated.

Given upper and lower bounds on ϵ_j that are consistent with the observed n_j, the likelihood of model parameters is

$$L(n|\Theta) = \prod_{j=1}^{J} \int_{b_j^{\ell}(n_j)}^{b_j^u(n_j)} \frac{1}{\sigma} \phi\left(\frac{\epsilon_j}{\sigma}\right) d\epsilon_j = \prod_{j=1}^{J} \frac{1}{\sigma} \left\{ \Phi\left[\frac{b_j^u(n_j)}{\sigma} \right] - \Phi\left[\frac{b_j^{\ell}(n_j)}{\sigma} \right] \right\}, \tag{C.7}$$

where n denotes the vector of sample sizes for all experiments, Θ denotes the set of model parameters, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal p.d.f. and c.d.f. respectively.

Finally, the data-augmented posterior distribution of the model parameters is proportional to $L(n|\Theta)\,\pi(\Theta)$. Estimates of the model parameters are obtained by sampling from this distribution via the Hamiltonian MCMC algorithm implemented in CmdStan 2.8 (Stan Development Team 2015).

C.3 Approach to Counterfactual Simulations

Samples drawn from the posterior distribution of the model parameters form the basis for the counterfactual simulations. Given a set of model parameters, choice of sample size for each experiment is simulated by 1) sampling the random component of the experiment's payoff (ϵ_j) from a truncated normal distribution bounded by $b_j^{\ell}(n_j)$ and $b_j^u(n_j)$ (n_j being the observed sample size for study j), 2) calculating the net expected payoff of the experiment under various counterfactual sample sizes,

and 3) choosing the sample size with the greatest net expected payoff.

Repeating this exercise for each set of parameters sampled from the posterior distribution, and then taking the average over the simulated draws of n_j, produces a vector of expected sample sizes. This set of sample sizes represents the *baseline* case for the purpose of subsequent comparison.

For each simulated policy, the procedure is repeated after manipulating the model to reflect the conditions specified by the intervention. The output of this simulation is again a vector of predicted sample sizes, only now the predictions fall under the auspices of the *counterfactual* policy. Predicted sample sizes are then compared under the counterfactual and baseline policies.

D Expanding Lab Capacity

This section reports results for a third policy intervention not reported or discussed in the main text. This intervention targets the incentive to minimize the amount of time spent collecting data by doubling the lab's capacity (it is referred to as the "lab capacity" intervention). Doubling the lab's capacity has two effects: 1) it decreases by half the number of days needed to run an experiment with its original sample size, and 2) it cuts in half the number of other studies using the lab on the same day (assuming proper coordination). Accordingly Equations (3) and (12) in the main text become:

$$C_j(n) = p_j n + \lambda_j \left[D_j\left(\tfrac{1}{2}n\right) \right]^{\delta} + F_j \tag{D.1}$$

$$\log(\lambda_j) = \log\left(f_\lambda(\mathcal{R}_j) \right) + \text{paid}_j \beta_{\lambda,\text{paid}} + \text{time}_j \beta_{\lambda,\text{time}} + \tfrac{1}{2}\text{other}_j \beta_{\lambda,\text{other}} \tag{D.2}$$

As with the interventions targeting reimbursement, this simulation assumes the supply of study participants is sufficient to satisfy any higher demand.

Results are presented in Table 5. Compared to the reimbursement policies, the lab capacity intervention would do very little to increase sample sizes, and would cost far more to implement. Doubling the capacity of the lab leads to an average increase of about 1%, although some experiments have predicted increases as high as 30–40%.

The subset of experiments with the greatest expected sample size increases (10% or more) share

Table 5 Model Comparison

	MODEL				
	Simple	*R*	*E*	*R+E*	*Full*
Source of heterogeneity					
Observed researcher characteristics ($\bar{\theta}_g$, γ_θ)		x		x	x
Observed experiment characteristics (β_θ)			x	x	x
Unobserved researcher characteristics (τ^2_θ)					x
RMSE of posterior predictions (%)	78.0	77.2	75.6	74.8	64.6
Average (%) sample size increase (policy, subject pool, subsidy)					
Cash, paid only, $s = 1/3$	26.8	24.6	27.1	23.8	23.6
Cash, both, $s = 1/3$	13.1	11.3	11.2	10.9	12.1
Credits, both, $s = 1/3$	5.3	7.3	9.1	9.6	10.2
Credits, credit only, $s = 1/3$	21.6	22.0	23.8	24.2	24.1
Lab, paid only	.1	.2	.2	1.0	.9
Lab, both	.2	.2	.3	.3	1.1
Lab, credit only	.1	.1	.2	.2	.7

a few characteristics: 1) they typically used the credit pool, 2) they had relatively small samples, 3) they took fewer days to run, and 4) they always scheduled participants in the final time slot on the last day of data collection. With such limited benefits, expanding lab capacity does not make sense for this institution. However there may be other reasons to expand the capacity of this lab, and other labs with lower capacity might find such an intervention to be more effective at increasing sample sizes.

E Results for other Model Specifications

Counterfactual results for all models are shown in Table 5. Results are qualitatively similar across all specifications. However, as the degree of heterogeneity increases, the model's ability to rationalize the end-of-day effect (whereby studies are scheduled in the last slot on the final day of data collection) improves. Hence the full model shows the highest estimated increases in sample size under the lab policy. By contrast, estimated improvements under the cash intervention are somewhat lower, and the estimated improvements under the credit intervention are somewhat higher, for models with more heterogeneity.