Consumers of Experimental Observations: Understanding How Experimental Costs Affect Sample Size and Composition

Jason M.T. Roos*

13 August 2018

Abstract

Experimental samples that are too small to detect true effects have plagued the behavioral sciences for years. At best, these so-called underpowered studies provide weak evidence for measured effects. At worst, they add false positive results to the literature and are a waste of limited research money. Obtaining bigger samples is expensive, and yet previous empirical research has not considered the role of experimental costs. This study shows how much costs matter by analyzing a novel data set describing all experiments at a behavioral lab over many years. Demand for paid participants at this lab is inelastic: a 10% percent higher reimbursement rate corresponds with a 6% smaller sample. Studies reimbursing students with course credit rather than money have significantly larger samples, but incentivizing students with credit creates a subtle and important selection problem. These and other results show how researchers conduct science with limited resources, leading to new insights relevant to how we fund, incentivize, and reform behavioral research.

*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. Statement of interest: The lab whose data are used in this study is partially funded by the same organization that funds the author's research. This relationship has not led to any direct or indirect conflicts of interest.

Consumers of Experimental Observations: Understanding How Experimental Costs Affect Sample Size and Composition

Imagine that you have planned an experiment using paid participants, and that it will cost $800 to run. You then run a small pilot study and learn that the manipulation takes more time than expected. Apart from taking longer to run the experiment's design is exactly the same, only now it costs $1000 instead of $800. Would you decrease the sample size from what you planned? If not, can you think of a researcher who might?

Scientists have finite resources with which to carry out their work. For experimental or survey-based research, this constraint means researchers face two competing incentives. One is the need to gain knowledge about a population of interest with the greatest precision. The other is the need to minimize the cost of obtaining that knowledge. For many in the behavioral sciences, the main drivers of a study's cost are the amount of money paid to each participant and the total sample size. Holding fixed the study's design, collecting a larger sample leads to greater precision. But a larger sample also means a higher cost. So much in the same way the amount of a good purchased reflects consumers' trade-offs between higher consumption and lower spending, sample sizes reflect researchers' trade-offs between higher precision and lower cost (Blattberg, 1979; Cohen, 1992b; Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997; Gelman & Carlin, 2014).

The idea that the cost of running an experiment can affect its sample size and composition is not new. At the same time, the extent to which costs affect experimental samples has never been quantified. The real-world consequences of this trade-off are unclear, yet it is crucial for us to understand them. The need to balance better science with lower costs affects not only researchers, but also the organizations facilitating and funding their work.

Experimental costs affect the way researchers conduct science in ways that are both obvious and subtle. The goal of this paper is to understand and raise awareness about these effects. New insights emerge from analysis of a novel data set describing all behavioral experiments at a behavioral lab over many years. Underpinning the analysis is the idea that researchers are con-

sumers of the information generated by study participants. Using ideas and tools from consumer research leads to new insights about how researchers conduct their work, and these insights bear directly on the issues of how to fund and incentivize better science.

Why Sample Sizes Matter

The empirical analysis in this study focuses on the drivers of sample sizes in lab-based behavioral experiments. Because sample sizes are directly related to the amount of information an experiment generates, there are many reasons to care about what determines them.

For one, small samples provide, at best, weak evidence in support of a measured effect. All else equal, the smaller the sample, the more variable the measured effects. Hence, the smaller the sample, the lower the experiment's power (defined as one minus the expected false negative, or type II error rate). Small-sample variability and low power raise the chance that published results are in error (Cohen, 1992a; Kraemer, Gardner, Brooks III, & Yesavage, 1998; Maxwell, 2004). From the standpoint of generating new knowledge, larger samples should generally be preferred. At the same time, small samples do not always imply low power in absolute terms. Importantly, there are many ways to increase power beyond obtaining bigger samples (Allison et al., 1997; McClelland, 2000; Maxwell, 2004; Abraham & Russell, 2008; Button et al., 2013; Meyvis & Van Osselaer, 2017). But after an experiment's design has been set, a bigger sample is the only remaining tool for increasing precision.

Another reason to care about sample sizes is money. Running experiments with almost no chance to measure a true effect is potentially a waste of limited resources (Halpern, Karlawish, & Berlin, 2002). Behavioral experiments take time and money to carry out. Still, researchers often apply resources to studies with samples that are too small to measure anything useful (Ioannidis, 2005). To limit such waste we need to know how costs affect researchers' decisions.

Yet another reason to care about sample sizes is related to replicability and the movement to reform research practices. Many proposed reforms would place tighter controls on false positive (type I) error rates. There is increasing recognition that one consequence of these reforms will

be the need for bigger samples (Maxwell, 2004; Ioannidis, 2005; Shen et al., 2011; J. P. Simmons, Nelson, & Simonsohn, 2011; Bakker, van Dijk, & Wicherts, 2012; Schimmack, 2012; Fiedler, Kutzner, & Krueger, 2012; Asendorpf et al., 2013; Button et al., 2013; J. P. Simmons, Nelson, & Simonsohn, 2013; Miguel et al., 2014; J. Simmons, 2014; Inman, Campbell, Kirmani, & Price, 2018). Researchers might soon find themselves in need of bigger samples, while lacking the resources to collect them. (Maxwell, 2004; Kahn, 2007; Baumeister, 2016). It matters whether limits on researchers' resources will get in the way of better science, and it matters whether funding organizations can deploy resources to remove those barriers.

Researchers as Consumers of Information

Previous work has considered researchers in their role as *producers* of new knowledge (Stephan, 1996; Dasgupta & David, 1994). This paper takes a different approach and instead considers researchers in their role as *consumers* of information. To illustrate the value of this framing, imagine a two-cell, between-subjects experiment. The study is ready to schedule in the lab, and one hundred participants have signed up. But there is still room (and enough participants in the pool) to schedule 10 more. Each participant will receive $5, so adding another 10 increases the study's cost by $50. Adding 10 observations also raises the experiment's power from 76% to 80%. Is this 4% increase in power worth an extra $50? A researcher with limited funds might not think so, whereas others might disagree.

This choice mirrors the type of trade-off consumers make all the time. But in contrast to other settings where we study consumption choices, we hold researchers to the normative standard that they they should not be cost-sensitive when choosing their sample sizes. Rather, we expect them to choose sample sizes that exceed some minimum threshold of statistical power (typically .80 for a type I rate less than .05). The standard tools we have developed for researchers to choose sample sizes, such as G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), do not account for the cost of obtaining the sample. In light of what we know about consumer choice in other domains, it should come as no surprise that experiments powered greater than .80 are more the exception

than the rule.

Prior Work and Contribution

This paper contributes to research seeking to understand why low-powered experiments have persisted so long in the behavioral sciences. Modern work in this area was spurred by Cohen's (1962) meta-analysis of articles from the *Journal of Abnormal Psychology,* in which (post hoc) median power was found to be .17, .46, and .89 for small, medium, and large effects. This finding led to the development of tools for conducting a priori power analysis, and it helped spread awareness about the benefits of bigger samples.

A notable result of this early work was Cohen's (1969) proposal to choose sample sizes achieving power of at least .80. Although Cohen's 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 and not a strict requirement. But like other defaults, .80 has evolved into a normative standard (in spite of being "arbitrary").

Subsequent meta-analyses of published experiments suggest that experimental power has not increased much—if at all—since Cohen's seminal paper (Sedlmeier & Gigerenzer, 1989; Maxwell, 2004; Shen et al., 2011; Marszalek, Barber, Kohlhart, & Holmes, 2011). One reason why .80 is normatively expected, but often ignored, is that a standard power analysis does not account for experimental costs.

Cohen did acknowledge that powering an experiment at .80 might not be feasible if costs are too high; he noted that power analysis then "leads to the useful discovery that the research as

conceived is not viable" (Cohen, 1992b, p. 100). But rather than abandoning their experiments, many researchers proceed anyway, choosing sample sizes in light of their available resources (Maxwell, 2004). This behavior bears some resemblance to consumers with limited budgets, who often settle for lower quality goods rather than purchasing nothing at all.

Perhaps in recognition of the way researchers actually work, some studies in this literature explicitly address the trade-off between the costs and benefits of running an experiment. This work is theoretical and prescriptive, intended to help researchers make best use of limited resources (Blattberg, 1979; Ginter, Cooper, Obermiller, & Page Jr, 1981; Sawyer & Ball, 1981; Chatterjee, Eliashberg, Gatignon, & Lodish, 1988; Cohen, 1992a; Allison et al., 1997; Moscarini & Smith, 2002; Winkens, Schouten, van Breukelen, & Berger, 2006). Absent from the literature though are empirical studies examining researchers' choices. This paper helps to fill that gap by analyzing experimental metadata from a diverse group of behavioral researchers over a period of many years. The data themselves are unique to the literature, and they provide a much-needed view into how behavioral researchers conduct science on a day-to-day basis.

The data used for this study come from a lab participant scheduling system. Thus they include information about both published and unpublished experiments. Inclusion of the latter group is crucial for gaining a more complete understanding of factors that affect sample sizes. By necessity, previous empirical studies considered only sample sizes reported in published articles. But it is well established that this leads to biased conclusions, through what is known as *publication bias* or the *file drawer problem* (Sterling, 1959; Rosenthal, 1979). The analysis in this paper does not suffer from these limitations.

The analysis generates new insights about how resource limits affect researchers' behaviors. They reveal intriguing relationships between sample sizes and the way participants are compensated (either with money or course credit), how much participants are paid, and even the time of year they are recruited. As one might expect, studies paying participants with money have smaller sample sizes than those using (free) students, and studies requiring more money per participant tend to have the smallest samples. But understanding how much these effects stem from

the use of money, rather than from other factors correlated with the use of money, requires careful analysis.

Thus, one novel contribution stemming from this analysis is to measure for the first time researchers' price elasticity of demand for paid participants. Demand for study participants at this lab is inelastic, on the order of about $-.6$. This means that, all else equal, a difference of 10% in the cost of paid participants corresponds with about a 6% smaller sample.

Other results from this analysis suggest researchers behave in ways the help counteract the effects of higher costs. One example is that studies conducted by a single researcher use the fewest number of paid participants, whereas studies conducted by two researchers use twice as many. This points to researcher collaborations as a potentially powerful tool for improving the quality of experimental research.

Other findings pertain to important differences in the composition of samples paid for with course credit or money. Experiments compensating students with course credit sample from a population that differs at the start and end the academic term. Some students systematically participate in studies at the start of each academic terms, and others at the end. The result is a high degree of self-selection. This may not be evident to researchers, but may affect their experimental outcomes. By contrast, there is no evidence for a similar selection effect among participants receiving money.

In total, these results show how cost differences can affect the way behavioral researchers work. They bear directly on issues relevant to scientific inference and science reform, and they demonstrate the value of applying concepts and tools from consumer research in the study of researchers.

The remainder of this paper follows a sequence of analyses based on archival lab data. It begins with a discussion of the major features of the data set, and a high-level analysis of the distribution of sample sizes at this lab. A detailed descriptive analyses of experimental factors associated with sample sizes then follows. Building on this is a regression analysis that isolates the impact of reimbursement on the number of paid participants used and leads to an estimated elastic-

ity of demand for paid participants. The last part of the analysis considers the potential trade-off between bigger samples and fewer studies. The paper then concludes with a general discussion of the results and their limits.

Archival Data from a Behavioral Lab

The data used for this study come 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 archive includes metadata for 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 19, 2014. These experiments were performed by a diverse group of behavioral scientists at all academic ranks (including graduate students). Most researchers in the data 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 are affiliated with RSM (and in particular, the Department of Marketing Management).

Overview of the Data Set

The archive contains metadata describing the experiments, researchers, and participants involved in the lab. The archive is organized by whether participants are reimbursed with course credit (the *credit pool*) or money (the *paid pool*). Titles, study descriptions, and lists of researchers associated with each study identify a subset that used participants drawn from both pools. The focus of this paper is on studies executed in the lab premises, hence a small number of studies for which data collection took place outside the behavioral lab (e.g., at the nearby Erasmus Medical Center) 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 final data set used for the analysis describes 809 experiments associated with 134 researchers.

Table 1. Descriptive Statistics

	Min	25%	Median	Mean	75%	Max
Experiments ($N = 809$)						
Duration, minutes	15	30	30	39.6	45	270
Time slots	1	74	160	193.0	276	1445
Sample size	1	36	75	93.5	132	582
Associated researchers	1	1	1	1.5	2	7
Researchers ($N = 134$)						
Experiments	1	1	3	7.4	10	102
Distinct collaborators	0	0	1	2.4	3	18
Days with data collection	1	7	19	34.9	45	197
Months in system	<1	1	10	17.4	31	80
Participants ($N = 11{,}995$)						
Total experiments	1	2	4	6.3	9	99
Months in system	<1	<1	7	11.8	20	72

The archival data identify many things—who participated in each study, when the study took place, how much time was needed from each participant, the type and amount of compensation participants received, and which researchers were involved with the study. But they do not contain details about study designs, nor any of the data generated by the experiments themselves. Summary statistics for experiments, researchers, and participants are summarized in table 1, and are discussed below.

Experiments and Time Slots. Registering an experiment in the scheduling system entails entering basic details relevant to the study. This includes information such as a title and description, and any other researchers involved. Researchers also indicate the time needed to collect data from each participant (the study's *duration*). Most studies last 30 or 60 minutes. The data also record the type and amount of compensation. Fifty-nine percent of studies compensated participants exclusively with course credit, 26% exclusively with money, and 15% using a mix of the two (meaning some students received money, others course credit). Study duration and participant payment are closely related to sample sizes, and these relationships are considered in detail later.

After recording basic information about the study, *time slots* are *opened*. An open time slot specifies both the timing and the facilities (e.g. rooms and their equipment) needed to interact

with one or more participants. The chosen facilities establish the maximum number of participants who can sign up for the same time slot. For example, a typical room with cubicles and computers can handle between four and eight participants at once. Notably, this lab has much higher capacity than labs at similar institutions. The lab occupies over 6,000 square feet, and the ten busiest days in the archive saw more than 275 participants each.

The number of participants who eventually sign up for a study—that is, the number of time slots *filled* by participants—determines the study's sample size. There are typically more time slots opened than filled. The relationship between opened and filled slots is also considered subsequently.

Apart from the availability of lab facilities, there are no constraints on how many time slots a researcher can open for each study. This means a researcher lacking the budget to pay for an experiment would not be prevented from running it, because budgets are reconciled *after* studies have concluded. Similarly, there is no formal allocation of course credit to researchers. In principle, any researcher can offer to compensate participants with course credit (strong norms for appropriate lab use help prevent abuse). An implication of these rules is that researchers budgets do not sharply bind for any given experiment. Any study could reasonably have included a few more observations.

Researchers. Each experiment is associated with one or more researchers. Among the studies conducted at this lab, 70% are associated with one researcher, 20% with two, and the remaining 10% with three or more. The median researcher is associated with only three studies. But 25% of researchers are associated with 12 or more studies. Half of the researchers used the paid pool for about half their studies. But there are researchers who never use the paid pool at all. The data contain no information about researchers apart from their email addresses. These identify the researcher's university division, or whether the researcher is a graduate student. Average sample sizes vary considerably with the number of researchers involved in each study and their university divisions.[1]

[1]The limited data describing university divisions provides a source of observed heterogeneity that improves statistical efficiency. But because results related to these data could be harmful to specific researchers, they are not

Participants. The archival data describe 75,636 observations collected from 11,995 unique study participants. Eighty-one percent of these individuals are students who participated in studies for course credit. The remainder are a mix of students and non-students who participated in studies in exchange for money. Five percent of students who earned course credit also participated in one or more paid studies. The median study participant took part in four different experiments, over seven months. Participants, especially students, maybe involved in many experiments. Thus researchers can encounter the same individual in multiple studies. Indeed, one person participated in 99 paid studies over the course of many years. Still, the median sample for the median researcher contains 98% participants who are never encountered in another experiment.

Distribution of Sample Sizes

The focal variable for this paper is sample size, which varies considerably across studies. Two approaches to characterizing this variation are presented next before considering how sample sizes relate to the other variables described above.

Sample Sizes Over Time. The first approach is to look at how sample sizes have evolved over time. In figure 1, each study is indicated by a point located along the *x*-axis on the day the first participants were scheduled, and along the *y*-axis at the study's final sample size. Strong seasonal patterns due to summer and winter holidays can be seen as regions in which no new studies began. Average sample sizes over time are shown by the LOESS regression line, which runs along the *x*-axis. This line is relatively flat, but there is some indication that sample sizes might have started increasing around the time of the 2011–12 academic year. A number of galvanizing events in the recent movement to reform scientific practice occurred around this time, including publication of J. P. Simmons et al. (2011) (Gelman, 2016).

Are the sample sizes shown in figure 1 large or small on average? The answer depends on the design of the experiments and the underlying research questions. An experiment after all can

---

reported numerically.

Figure 1. Sample Sizes over Time

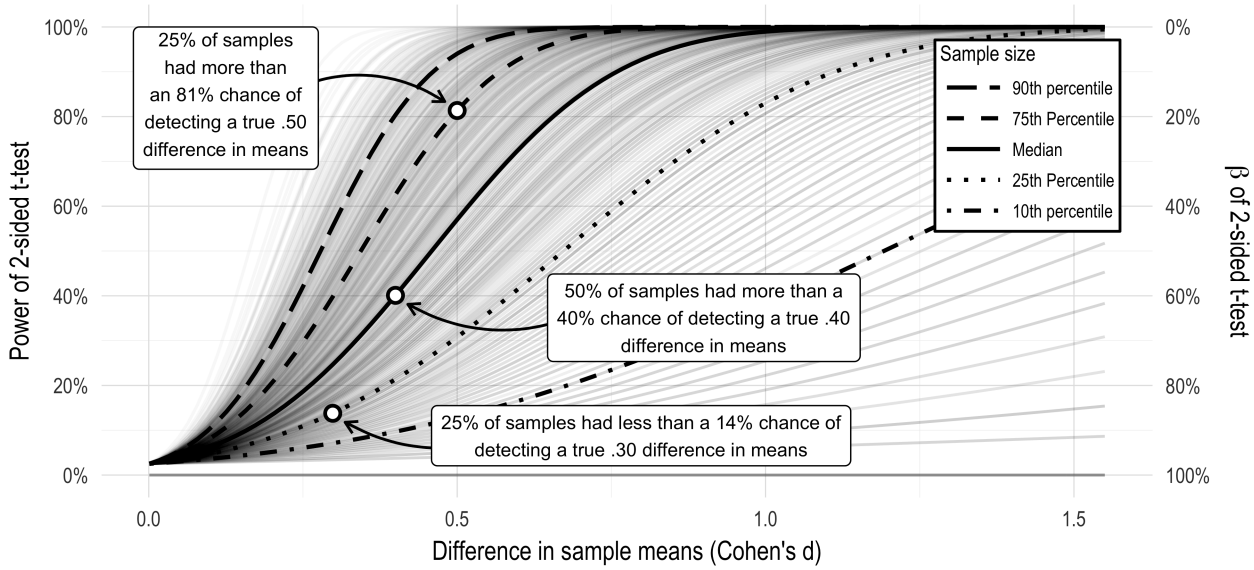
Notes. Each point represents an experiment. Empty vertical regions correspond with the summer and winter holidays. The LOESS regression line (with 95% CI in grey) shows how average sample sizes have changed over time.

achieve sufficient precision to generate new knowledge using a relatively small sample if, for example, it has a simple within-subjects design, uses a manipulation generating a large effect, makes repeated measures, uses validated instruments, has a preregistered protocol, etc. By the same token, an experiment might have low precision in spite of a large sample if it uses, for example, a complex between-subjects design, relies on weak manipulations, has a noisy dependent measure, is unfocussed in its objectives, etc. There are certainly examples of both high- and low-precision studies in figure 1.

A Whole-Lab Power Curve. The second approach to characterizing sample sizes seeks to quantify the precision that might be obtained from a *typical* sample collected at this lab. Accordingly, figure 2 plots the full distribution of sample sizes under the assumption that each sample could have been used to generate inferences from a two-cell, between-subjects experiment, using a Student's *t*-test with $\alpha = .05$. In this *whole-lab power curve*, each sample is represented by a line connecting standardized effect sizes (Cohen's *d*) with the power (or type II error rate) of a hypothetical *t*-test. This whole-lab power curve offers a glimpse into the types of effects that might be detected by typical samples at this lab.

To be clear: the whole-lab power curve is a vast simplification of the complexity entailed in

Figure 2. A Whole-Lab Power Curve



Notes. Each line represents a sample collected at this lab, and describes the attainable effect size and power for a hypothetical $\alpha = .05$ Student's *t*-test based on that sample.

conducting behavioral science. Figure 2 contains a wide range of studies, including many that were "successful," "unsuccessful," pre-tests, hypothesis tests, exploratory, and confirmatory. In reality, many studies at this lab would have achieved far greater precision than this hypothetical *t*-test baseline, and others worse. For some studies, the concept of statistical power would not even apply.

The value of figure 2 is that it facilitates comparison of raw sample sizes from an *uncensored* set of studies conducted at a single lab. And it makes this comparison using a common analytical framework for which many behavioral researchers have strong intuitions. The whole-lab power curve can be a useful diagnostic tool for understanding behavior within a single lab. It can also be used to compare behavior across labs or groups of researchers.

For example, consider the median and outer quartile sample sizes, which are highlighted in figure 2. The power curves for these quartiles show 25% of the samples collected at this lab would have had over 80% power for a *t*-test detecting an effect size of $d = .5$ or larger. At the same time, 25% would have had less than 14% power for effects as small as $d = .3$. By translating sample sizes into hypothetical *t*-tests, the implications of their distribution can be understood

more easily than when considering raw sample sizes alone.

Factors Affecting Sample Size and Composition

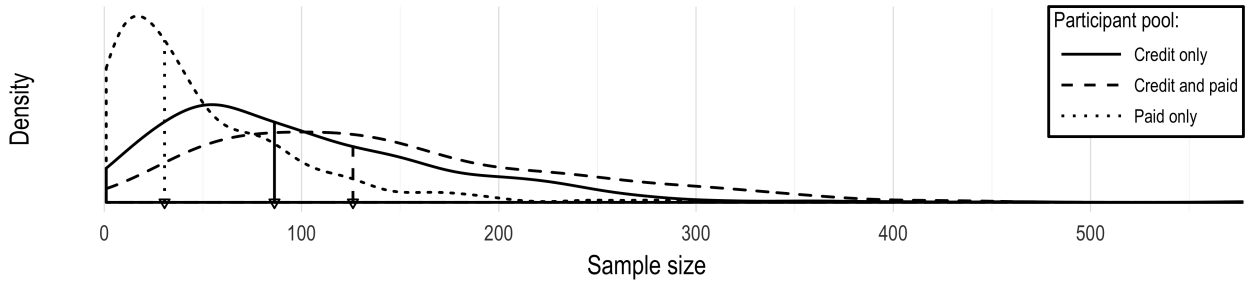
There are many factors influencing sample sizes that lie within and outside of researchers' control. The sample sizes shown in figures 1 and 2 thus represent final outcomes after a series of researchers' and participants' decisions. All of these decisions occurred in the context of a dynamic lab environment. The analysis in this section considers some of these factors, including: 1) which pool the participants are recruited from, and by extension whether they are reimbursed with money or course credit; 2) the number of researchers associated with each study; 3) the time needed to collect data from each participant, and by extension, the amount of money or course credit needed to compensate them; and 4) the availability of research participants over the course of the academic calendar.

Each of these variables relates to final sample sizes (and the composition of those samples) in ways that shed light on how researchers make use of limited resources. These relationships are interesting in their own right, and are therefore explored in detail in the analysis that follows. These results also motivate and lay a foundation for the more integrative regression analysis that follows.

Sample Sizes Differ Between the Paid and Credit Pools

Figure 3 shows the distribution of sample sizes while grouping studies together by how they recruited and reimbursed participants: 1) exclusively from the credit pool, 2) exclusively from the paid pool, or 3) using a mix of participants recruited from both pools. There are striking differences among the three distributions. Studies conducted exclusively with paid participants have much smaller sample sizes (median 31) compared to studies using course credit for all (86) or some (126) participants.

Figure 3. Distribution of Sample Sizes by Participant Pool



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

Experiments Using Exclusively Paid Participants Have the Smallest Samples. Figure 3 shows that studies using exclusively paid participants have considerably smaller samples than studies using course credit. There are many reasons why sample sizes might differ so much between participant pools. For one, compensating participants with money allows the use of incentive-compatible outcomes. By focussing participants' attention on the task at hand, incentive-aligned tasks can lead to more precise measurements, which in turn permits a smaller sample.

Another reason could be related to task duration. In the credit pool, almost all studies last 30 or 60 minutes. By contrast, the paid pool includes many studies requiring fewer than 30 minutes to administer. The paid pool might be better suited for running quick, small-sample pre-tests and pilot studies.

And of course, money is another reason why samples might be smaller in the paid pool. Compensating participants at the standard rate of €10 per hour makes paid studies, all else equal, far more expensive than studies using course credit. Researchers do not have unlimited budgets and this helps explain why samples using paid participants are smaller.

Experiments Using Participants from Both Pools Have Larger Samples. Recall from figure 3 that the subset of studies with the biggest sample sizes are those using a mix of participants from the credit and paid pools. If money is such an important, limiting factor when using paid participants, why aren't sample sizes in this group *smaller* than those conducted exclusively for course credit?

An explanation suggested by researchers using this lab is that studies using both course credit and money begin as studies intended to be run exclusively in the credit pool. If a study fails to attract sufficiently many participants from the credit pool—when consumers of information, in a sense, are forced to cope with a stockout—researchers have a choice. They can accept a smaller sample than they originally hoped for, or they can register the experiment in the paid pool in order to attract additional (paid) participants.

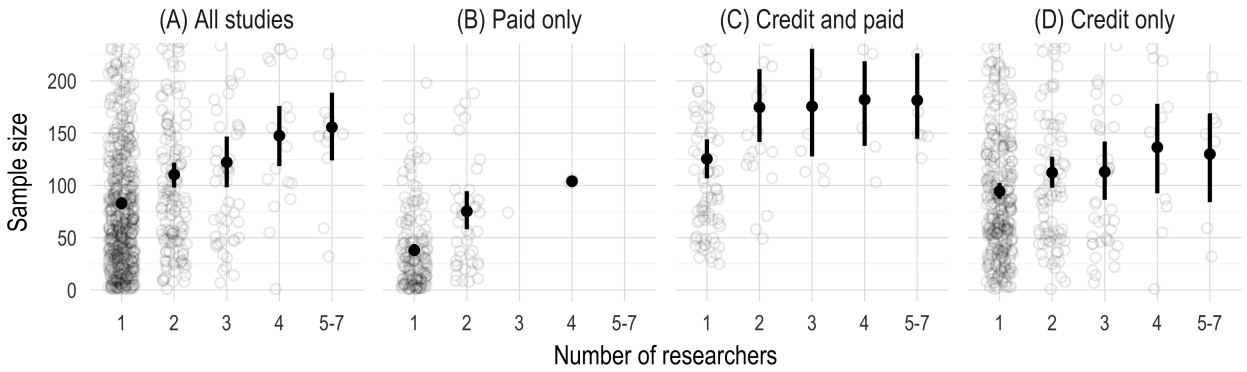
If correct, this explanation can account for why studies using both pools have larger samples on average. Just as the least price conscious consumers would be most likely to switch to a more expensive brand in the event of a stockout, it may be that the least cost-conscious researchers are most likely to switch their studies to the paid pool. It would follow that the studies using both course credit and money tend to have the biggest samples, as they are conducted by the least cost-sensitive researchers.

The data can potentially contradict this explanation. First, this explanation implies that studies using both money and credit typically collect data from credit pool first. This is in fact the case. Among studies using both pools, the first paid participant was scheduled on average two days after the first participant from the credit pool. Studies using participants from both pools do indeed start out in the credit pool.

Second, if researchers using both pools for the same study are less cost sensitive on average, their relative cost insensitivity should be evident from looking at their paid-only studies. Indeed, among researchers who used both pools, the average sample size for paid-only studies is 63.4 (SE 5.4). But for all other researchers, the average paid study has a sample size of 37.0 (3.8).

This evidence supports the idea that the paid pool is sometimes used by relatively less cost-sensitive researchers in order to make up for unexpectedly low recruitment rates among credit-reimbursed students. It also illustrates another idea: researchers want their studies to conclude as quickly as possible. In theory, a researcher can continue opening slots for weeks or even months until the desired sample size is attained. In practice this rarely happens. Rather, researchers abandon these studies or open slots in the paid pool. The data unfortunately do not describe researchers'

Figure 4. Sample Sizes by Team Size and Type of Compensation



Notes. Points indicate group means and bars their 95% bootstrap intervals. Panel (A) shows all 809 studies, panels (B)–(D) show subsets depending on whether participants were compensated with money, course credit, or both. Experiments with sample sizes greater than 225 are used to calculate group means and bootstrap intervals, but are not shown.

use of online panels. But the desire to collect data as quickly as possible may be a reason that services like Amazon Mechanical Turk (MTurk) are so popular with behavioral researchers.

Larger Teams Have Larger Sample Sizes

Sample sizes also vary with the number of researchers associated with each study. The median sample size is 64 (mean 82) among the 70% of experiments conducted by a single researcher, 92 (110) among the 20% associated with two researchers, 113 (122) among the 6.3% with three, and 149 (151) among the 4% associated with four to seven. Group averages and 95% bootstrap confidence intervals for all studies are shown in figure 4A.

The relationship between the number of researchers involved in an experiment and its final sample size is positive when considering all studies. But it is the subset of studies using paid participants that contributes the most to this pattern. This can be seen by comparing figures 4B and 4C, which show studies conducted in the paid pool, against figure 4D, which shows studies conducted exclusively in the credit pool. Indeed, studies using exclusively paid participants and linked to one researcher have a median (mean) sample size of 28 (38), whereas studies linked to two researchers have 65 (75).

The nature of these multi-researcher collaborations is not observed in the data. In many cases,

one or more collaborating researchers might be a graduate student (or research assistant). In other cases, collaborators may combine multiple short, unrelated experiments into a single session. For example, studies with four to seven researchers last 56 (SE 1.9) minutes on average, whereas those with one to three researchers last 39 (.7) minutes on average.

Another reason why larger teams could be associated with bigger samples is that when involving more researchers may produce a larger pool of money for paying participants (Kahn, 2007). An alternative explanation could be that studies that need bigger samples also need a greater number of junior researchers to manage data collection. But this is not supported by the data, as greater numbers of non-faculty (i.e., student) researchers are instead associated with studies with relatively smaller samples.

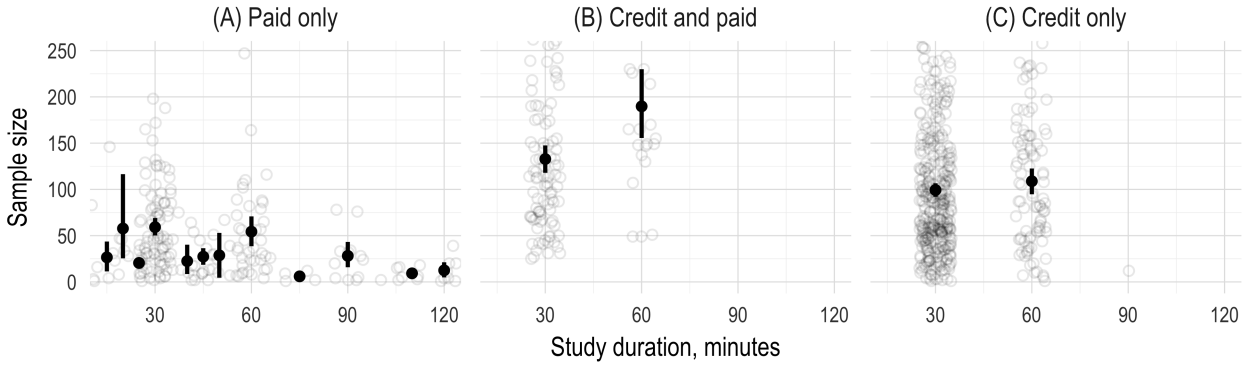
Duration and Sample Sizes in the Paid Pool

Turning next to the amount of time required to collect data from participants, figure 5 shows average sample sizes and 95% bootstrap CI's for studies of different durations. Once again, studies are grouped together based on how they compensated participants. The relationship between duration and sample size is clearly different among each group. In particular, sample size is negatively associated with study duration among studies using the paid pool only. This is not the case among studies using the credit pool.

There are many reasons for why this might be. One is that study duration is closely related to the overall experimental design. For example, a within-subjects design using repeated measures might take longer to administer, while attaining high precision with a relatively smaller sample. By the same token, a study requiring participant naïveté might permit just a single valid measurement per participant, thus taking less time to administer, while requiring a relatively larger sample.

Another explanation is that study duration is closely linked with the amount of money or course credit participants receive. The norm in this lab is to pay participants at a rate of two course credits or €10 per hour. All but one study using credit adhered to this norm. In the paid

Figure 5. Sample Sizes by Study Duration and Type of Compensation



Notes. Dark points indicate group means, and bars their 95% bootstrap intervals. Studies longer than two hours or with sample sizes greater than 260 are used to compute group means and bootstrap intervals, but are not displayed. Grey open circles indicate individual studies and are plotted horizontally jittered.

pool, 75% of studies followed the norm, 16% paid less at a median rate of €8/hr, and 9% paid more at a median rate of €12/hr. This explanation is supported by the pattern in figure 5A, which shows that among studies using only paid participants, those needing less time and money had larger sample sizes. By contrast, figures 5B and 5C show that among studies using course credit, longer studies generally had larger samples.

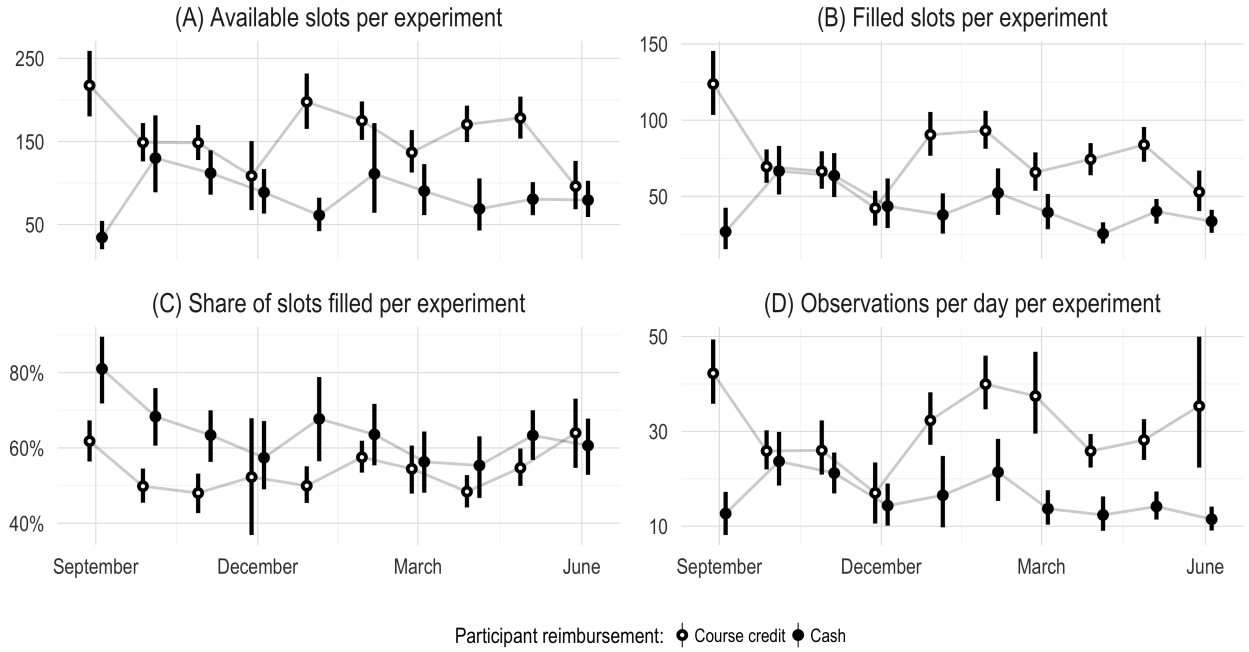
The pattern in figure 5A (showing studies run exclusively for money) is consistent with money costs affecting sample sizes. But the pattern in figure 5B (showing studies using a mix of money and course credit) is less clear. If researchers who conducted studies using both participant pools—as speculated earlier—are among the least cost-sensitive, then this could also explain the pattern in figure 5B.

Sample Sizes Vary Across the Academic Year

As previously noted, experiments almost never fill all of their open time slots. Researchers using this lab understand this and therefore open more slots than their target sample size. As figure 6 shows, the number of open and filled slots are highly variable over the course of the academic year.

First, consider use of the credit pool, which is indicated by the open circles in figure 6. Fig-

Figure 6. Supply and Demand for Participants by Month



Notes. Statistics are first tabulated for each experiment in each month, and then averaged across experiments. Experiments using both paid and credit-reimbursed participants are shown under whichever pool the data were collected from that month (this may include, for the same experiment, both participant pools).

ure 6A depicts the average number of slots that were opened, and figure 6B the number filled, among experiments using the lab each month. In September, at the start of the academic year, researchers make large numbers of slots available to student participants in need of course credit. These credit-reimbursed studies are able to fill a relatively large proportion of their open slots—about 60%, as shown in figure 6C. From this starting point in September, the numbers of opened and filled slots decrease over the next few months, before increasing again at the start of a new academic term in January (when students have new requirements for obtaining credit). At the end of the academic year in June, there are relatively few open slots available, and participants fill them at the highest rate.

These differences in the number of open time slots, as well as the rate at which they are filled, correspond with meaningful differences in the number of observations recorded each month and the number of lab days needed to collect them. Experiments using course credit collect an average of 124 observations in September, at a rate of 42 participants per lab day (figure 6D). But in

December, they collect an average of just 42 observations at a rate of 17 participants per lab day.

The situation is rather different in the paid pool. Researchers open fewer slots per experiment (figure 6A), but these open slots fill at a higher rate (figure 6C). The paid pool therefore provides a useful tool for researchers seeking to minimize the number of lab days needed—or, as previously discussed—trying to make up for lower than expected participation from the credit pool.

Student Samples Vary Across the Academic Year

Students who sign up for credit-reimbursed studies at the start of an academic term are not the same as those who sign up at the end. A student who signs up for a study in September is more likely to sign up for studies in January and April (at the start of subsequent academic terms). They are less likely to sign up in November, March, and June (at the end of those same terms). Conversely, many students consistently participate in credit-reimbursed studies at the end of each term, but not at the start. The correlation between the number of credit-reimbursed studies students participate in the first and last month of each term is −.14 (95% CI: −.16, −.12), whereas for paid studies, the same value is .60 (.58, .63). Appendix B reports correlations at the monthly level and describes corroborating results from a confirmatory cluster analysis.

These differences in study participation are a source of self-selection. Thus they can potentially affect the quality and generalizability of results obtained from credit-reimbursed experiments. Whether this selection is consequential or not depends entirely on the particulars of any given study and the underlying reasons why students sign up for studies early or late in the term.

To illustrate the potential bias arising from this behavior, consider the following. Assume that students signing up in September are typically more academically motivated and better organized than their counterparts signing up in November. An experiment trying to measure an effect that is also related to motivation or organization might have an easier or more difficult time measuring this effect at different points in the academic term. Worse, if motivation and organization affects behavior in one experimental condition but not the others, then conclusions drawn from the experiment will be biased. Moreover, the degree of bias will vary across the academic term.

Discussion

Taken together, these results show how differences in researchers' sensitivity to time and money costs can affect sample size and composition. Studies reimbursing participants exclusively with money have smaller samples on average than those reimbursing using course credit. Among paid studies, the longer (more expensive) ones have smaller samples than the shorter (less expensive) ones. Studies involving more researchers, which potentially draw on a larger pool of resources, use the greatest number of paid participants. Even the availability of student participants at different points of the academic year (and researchers willingness to wait for participants to sign up) play a role in determining the size and composition of experimental samples.

Are these differences due only to principled choices pertaining exclusively to the scientific content of the experiments? Or are they also influenced by other factors, such as limits on researchers' time and money? If these other factors indeed influence what these experiments eventually look like, are the researchers involved aware of how these forces affect their decisions? Most likely the answer is that some researchers are strongly impacted by these factors, whereas others are not, and that some are aware of this, while others are not.

Many of the patterns shown in this descriptive analysis can also be explained without appealing to researchers' cost sensitivity. Conversely, a single-minded aversion to higher costs cannot fully explain all of the patterns. Almost certainly, sample sizes are multiply determined by a range of factors, with money and non-money factors both playing important roles. The goal of the analysis that follows is to carefully isolate the role of money from these other factors.

Price Elasticity of Demand for Paid Participants

The analysis in this section seeks to measure the extent of researchers' sensitivity to higher costs when working with paid participants. The approach is to regress the number of paid participants in each study on the amount of money paid. This then leads to an estimate of researchers' price elasticity of demand.

Quantifying how reimbursement affects sample sizes is important because it is a first step to understanding how changes in lab or university policies might impact scientific outcomes. To give one such example, the standard hourly rate of reimbursement for paid participants has remained constant at this lab since its inception (this is in line with the median 1.9% annual rate of inflation during the period covered by the data; International Monetary Fund, 2015). Eventually the hourly rate will need to increase. When it does, how will researchers respond? Will increasing the expected rate of reimbursement without making any other changes decrease sample sizes? If yes, will the decrease be substantial? If the impact is substantial, how much would budgets need to increase to counteract the decrease? The answers to these questions depend on understanding the *extent* to which researchers are willing to trade off higher costs for better precision.

Method

The average price elasticity of demand is estimated from the following linear model.

$$\log\left(n_i\right) = \alpha + \log\left(p_i\right)\beta_p + c_i\beta_c + r_i\gamma_r + d_i\gamma_d + x_i'\phi + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right) \tag{1}$$

The variables describing each study i are:

- $\log n_i$, the (log-transformed) number of participants who are reimbursed with cash;
- $\log p_i$, the (log-transformed) amount paid to each participant;
- c_i, a variable indicating whether participants from the credit pool are also used;
- r_i, the number of researchers associated with the study;
- d_i, the difference between the study's duration and the average duration of all other studies by researchers associated with study i (first averaged across studies within each researcher, then across researchers);[2] and

[2] In a small number of cases where the set of other studies is empty, $d_i = 0$.

x_i, a vector of dummy variables indicating which groups the associated researchers are affiliated with.

The log-log form of the linear regression in (1) means the coefficient β_p has a straightforward interpretation as the price elasticity of demand for paid participants (i.e., the expected percent difference in sample sizes at payment levels differing by 1%). Studies using the credit pool exclusively are not included in the regression, as they have no paid participants ($n_i = 0$) and an undefined value for the reimbursement rate (p_i).

Estimates of the regression coefficients are obtained via instrumental variables (IV) regression. An IV approach is used because two of the variables—the money paid to each participant (p_i) and whether the credit pool is also used (c_i)—could be correlated with unobserved factors that also affect sample sizes. Such a correlation might occur, for example, if a researcher were to move a credit-only study to the paid pool, or offer a higher rate of payment in order to collect a bigger sample more quickly (as discussed in the previous section). If such a correlation exists, not accounting for it will lead to biased estimates of β_p and β_c.

The instrumental variables in the IV regression are: 1) the timing of the study, specifically a) the number of months since the start of the current term, b) the number of months until the end of the current term, and c) an interaction between these two variables; and 2) the average duration of all other studies conducted by the focal study's researchers. These variables should be correlated with the amount paid to participants and whether the credit pool was used, but uncorrelated with other unobserved factors affecting study i's sample size.

Four versions of the regression model are estimated. Models 1–3 are estimated from the $N = 332$ studies conducted exclusively or partially with paid participants. Model 4 is estimated from the $N = 209$ studies using only paid participants. Terms excluded from each of the different specifications are indicated by the empty cells in table 2.

Table 2. Regression of Log Sample Size on Payment Amount

		Model 1	Model 2		Model 3		Model 4	
Data	Parameter	OLS	OLS	IV	OLS	IV	OLS	IV
Intercept	α	2.02	2.16	3.23	2.43	3.30	3.01	3.67
		(0.33)	(0.43)	(0.69)	(0.46)	(0.66)	(0.72)	(0.48)
Log of money paid to each participant, log p_i	β_p		−0.06	−0.57	−0.19	−0.60	−0.31	−0.65
			(0.13)	(0.32)	(0.15)	(0.31)	(0.18)	(0.23)
Also used credit pool, c_i	β_c		0.03	−0.28	0.03	−0.26		
			(0.12)	(0.89)	(0.12)	(0.86)		
Number of researchers, r_i	γ_r	0.26	0.26	0.34	0.27	0.34	0.39	0.44
		(0.06)	(0.06)	(0.10)	(0.06)	(0.10)	(0.16)	(0.10)
Difference from average study duration, d_i	γ_d				0.34	0.59	0.37	0.59
					(0.20)	(0.27)	(0.24)	(0.28)
Dummy variables for researcher group(s), x_i	ϕ	Included	Included		Included		Included	
Residuals	σ	0.91	0.91	0.94	0.91	0.93	1.02	1.03
R^2		0.36	0.36	0.32	0.36	0.34	0.35	0.34
Experiments included		Any paid	Any paid		Any paid		Only paid	
N		332	332		332		209	

Notes. Dependent variable is the log of the number of participants reimbursed with money. Sandwich standard errors are reported for IV estimates.

Results

Estimates from the four regression models are shown in table 2. Diagnostic tests (reported in appendix C) support the appropriateness of IV regression and the chosen instruments. OLS estimates for all four models are also shown in table 2 for comparison.

Differences across the four models provide insights into how each of the regression variables explain differences in sample sizes. Consider model 1, which includes an intercept, researcher group dummies, and the number of researchers. This simple model explains about a third of the variation in the number of paid participants. Estimates of R^2 in the IV models are affected by the use of instruments, so an exact comparison between the OLS baseline model 1 and the IV models 2–4 is not directly informative. But it is clear that most of the variation in sample sizes is explained by the (faculty division) group dummies and number of researchers. The amount of variation explained by reimbursement rates is relatively small.

The estimated coefficient for $\log p_i$ is −.57 (SE .32) in model 2, −.60 (.31) in model 3, and

−.65 (.23) in model 4. Recall that due to the log-log form of this regression, these coefficient estimates can be interpreted as the price elasticity of demand for paid participants. Hence, a study with a 10% higher reimbursement rate has on average about a 6% smaller sample. The higher magnitude and smaller standard error of the estimated elasticity in model 4 suggests a higher degree of cost sensitivity among studies using participants recruited exclusively from the paid pool.

Models 2 and 3 include a dummy variable indicating whether the credit pool was also used (c_i). In both models, this coefficient is estimated with a negative sign but a relatively high standard error.

Finally, studies that are unusually long or short relative to what is typical for the researchers involved may require unusually large or small samples. Hence models 3 and 4 also include the difference between each study's duration and the average duration of all other studies conducted by that study's researchers (d_i). The value of this coefficient is .59 (SE .27) in model 3 and .59 (.28) in model 4. This is consistent with longer studies receiving bigger samples on average.

Discussion

Across all model specifications, the estimated price elasticity is negative with a magnitude less than one. This means researchers have *inelastic* demand for paid participants. The result implies two things. First, forcing a researcher to reimburse participants at a higher rate—say €11 per hour instead of the typical €10 (10% more)—would be expected, all else equal, to lead to about a 6% smaller sample size ($.10 \times -.60$ in model 3 or $.10 \times -.65$ in model 4). Second, because demand for paid participants is inelastic (i.e., $|-.6| < 1$), not all of the cost increase would be offset by the smaller sample. Instead, the researcher would bear the burden of a slightly higher total cost for the smaller experiment. To illustrate, an experiment with 100 participants receiving €10 each would cost €1000. Forcing a 10% increase in reimbursement to €11 would lower the sample size by 6% to 94 participants, but the experiment would now cost €1034 ($94 \times €11 = €1034$).

This interpretation is meant to convey the intuition behind this measurement. But it is an oversimplification of what would happen if the rate of compensation for all studies at this lab

were to change. Altering the amount researchers pay for participants would impact not only their sample sizes, but also which studies they choose to run. The full impact of this change would also depend on researchers' budgets, the willingness of participants to work for more money, and other factors. The IV regression controls for the most important of these factors. Still, the estimated elasticity should be interpreted as a first order approximation for how sample sizes among studies that made it to the lab might have been affected by higher or lower costs.

Although the rate of reimbursement matters a lot, another important finding from this analysis is that differences in the number of researchers involved and their faculty division (or student status) explain a lot. These variables account for much of the variation in the number of paid participants across studies. Such variation probably reflects differences in typical experimental paradigms across fields. A typical experiment in consumer behavior can be quite different from a typical experiment in experimental economics or cognitive psychology.

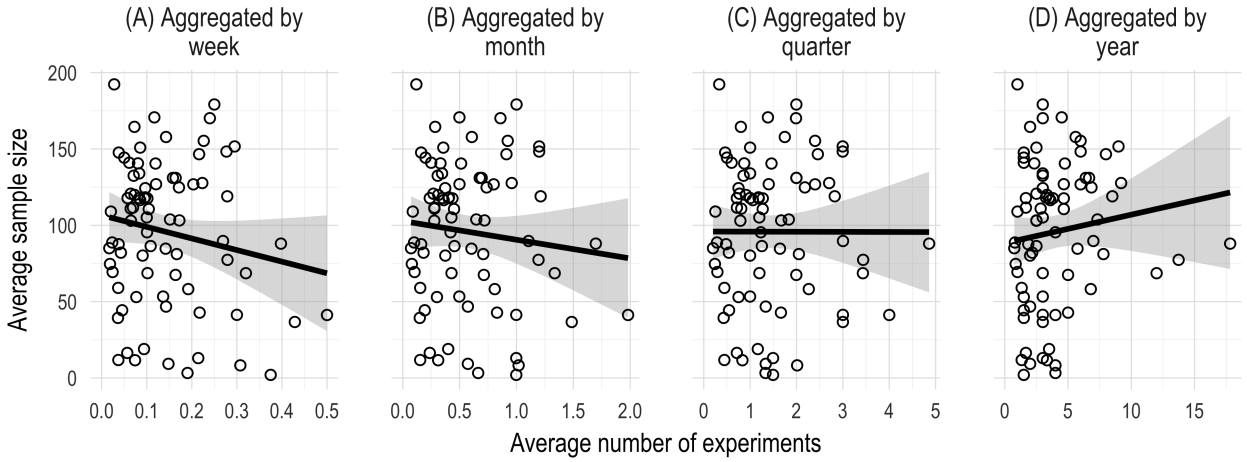
At the same time, differences in the way research is funded across academic subfields may also play a role. A typical research budget in consumer behavior can be quite different from a typical research budget in economics or psychology. Understanding the sources of these differences is beyond what these data can deliver. Budgets are not observed in the data after all. But gaining a better understanding for why these differences exist is necessary for deriving an efficient allocation of scientific resources at the institutional level (i.e., universities and scientific grantors).

Do Bigger Samples Imply Fewer Studies?

At any point in time, a researcher might have more than one study they could choose to conduct. But because researchers have limited time and money, the choice to run one study can come at the expense of running another. Does this mean there is a trade-off between running fewer experiments with bigger samples and more experiments with smaller samples, as has been suggested (Baumeister, 2016)?

Even though the data do not describe individual budgets, they can still shed light on this is-

Figure 7. Average Number of Studies and Experiments by Researcher over Different Time Frames



Notes. Points indicate averages for individual researchers who conducted two or more studies.

sue. If researchers must choose between smaller samples or fewer studies, then their average sample sizes would likely be (negatively) correlated with the number of experiments they run.

Figure 7 shows average behavior for each researcher associated with at least two studies. The *x*-axis indicates the average number of experiments conducted over four different time frames (shown in each of the four panels). The *y*-axis indicates the average sample size in those experiments. Figures 7A and 7B show averages for each researcher calculated at the weekly and monthly levels, respectively. These plots show a potentially (weak) negative association between running more experiments and running experiments with bigger samples over shorter time frames. At the weekly level, such a tradeoff makes sense because researchers have finite time and attention to focus on data collection. Figures 7C and 7D show the same averages, but calculated at the quarterly and yearly level. At longer time frames, there is no negative association between sample sizes and the number of experiments.

If researchers face a trade-off between larger samples and the number of studies conducted, in the short run it is more likely due to limited time than limited financial resources. If researchers face a trade-off on a longer time scale, it cannot be detected from these data.[3]

[3] An important caveat to this result is that evidence of a trade-off would need to be seen among the subset of studies actually making it to the lab. If the trade-off instead causes some studies not to be run at all (rather than having smaller samples), the analysis cannot detect it.

General Discussion

Differences in the resources needed to carry out experiments can affect the way researchers carry out their work. These effects are both meaningful and measurable. In particular, sample sizes at this lab vary between studies reimbursing with money and course credit. Studies relying on paid participants have far smaller samples than those using students, and higher reimbursement rates correspond with a smaller samples.

The idea that costs affect sample sizes is not new. Researchers working with paid participants are aware that the need to pay participants places limits on sample sizes. But the extent to which this affects researchers' decisions was previously unknown. Researchers at this lab are indeed sensitive to differences in the rate of reimbursement, with an estimated price elasticity of demand of $-.6$. This means that a 10% higher payment corresponds with a 6% smaller sample. Demand for paid participants is inelastic, though. This means that the smaller samples do not completely offset the higher reimbursement rate. Instead, researchers facing higher rates use smaller samples, but also face a higher total cost for the experiment. Consumer goods with inelastic demand are those that are most essential to the consumer. For experimental researchers, study participants are clearly essential goods.

Although researchers at this lab are cost sensitive, they do not avoid higher costs under all circumstances. Rather, the data also show how when the need for a large sample arises, researchers are willing to pay for it. The large samples found among studies started in the credit pool and later moved to the paid pool provide a good illustration of this. From the researcher's perspective, some studies are more important than others. Researchers prioritize where to deploy their resources in light of this. The analogy to consumer decision making applies here as well. Even the most price-sensitive consumer will buy a good at a high price if they need it enough.

Are researchers aware of the extent to which costs influence their work? Most consumers cannot articulate precisely how sensitive they are to small differences in retail prices. By the same token, most researchers would struggle to do the same for participant reimbursement. Like consumers, researchers can be aware that costs matter, while being unaware of exactly how much

costs affect their choices. But in contrast to consumers, we hold researchers to an impossible standard. We expect them to choose sample sizes based only on their expected type I and II error rates and effect sizes. Money isn't supposed to matter.

But researchers are people making choices in the face of limited resources. Of course money matters. Indeed, the data show evidence of behaviors that might strategically lower costs. Studies with larger teams use the most paid participants, and this might be because these teams have larger pools of resources to draw from. The magnitude of this association is considerable. Samples among paid studies run by two researchers are twice as big as those run by one researcher. As norms for the number of participants in a typical study rise, there may be more instances of co-authorship and other types of resource pooling. If these collaborations raise the quality of the (co-) authors' work, this should be viewed favorably by the field. An efficient use of limited resources should not be mistaken for a lack of academic independence.

Using student participants is yet another cost-lowering strategy. The marginal (money) cost from reimbursing each student participant is zero, so it should come as no surprise that credit-reimbursed studies have the largest samples at this lab. But relying on student participants has important implications for scientific outcomes. Many of these have been discussed in the literature (e.g., Peterson, 2001). The results from this study bring further clarity to some of these issues.

First, credit-reimbursed studies have samples for which participants have self-selected. Students are habitual about when they take part in studies, either at the start of end of an academic term. Selection bias can occur if the reasons behind when students participate are related to features of the experiment. Labs relying on student populations should be aware of this potential source of selection bias and attempt to counteract it. For example, rather than allowing students to sign up for studies, they can be randomized. Money-reimbursed studies at this lab do not suffer from this type of selection (although selection may arise from other sources not considered here).

Another point related to using students is that the number of available participants varies over

the academic year. When researchers cannot attract enough credit-reimbursed students, they may switch to paid participants. For many studies, mixing participants who are reimbursed differently does not affect inferences. But for other studies, pooling observations this way might bias a study's results. At the least, researchers should disclose the use of participants compensated in different ways. And they should be careful to control for differences in compensation during analysis.

Using credit-reimbursed students can be beneficial to science. Researchers can work with larger samples than they might otherwise have access to, and at a much lower cost. At the same time, requiring students to participate for free can introduce selection bias into studies. Paid participant pools (including online pools) may provide a more stable and diverse population. And money reimbursement can allow researchers to collect data more quickly. At the same time, using money to attract participants may lead to lower powered studies. There is an important trade-off here that researchers and their institutions should be aware of. They should seek to understand and limit the downsides of both types of reimbursement.

But apart from learning more about these phenomena in the context of their own labs, what else can researchers and institutions do? For one, individual researchers can be transparent about how they plan their work. They can document their decisions as they proceed and make these notes available to others. Benign choices seeking to limit experimental costs can look like bad science without proper context. Supplying that context allows researchers to do more with less.

Researchers can also stop using heuristics to choose sample sizes. Instead, they should try to derive sample sizes given not just their scientific goals, but also their resource constraints. The standard inputs used for power analysis (i.e., the expected effect size and acceptable error rates) are not enough. The expected costs of running a study also matter and should be an explicit part of the planning process. This will help researchers understand whether their studies need to be redesigned or abandoned.

In total, these results show us how resource-constrained researchers conduct behavioral science. They have implications bearing directly on issues of science funding and reform. Researchers'

choices often deviate from normative standards for how science should carried out. When this happens, it often raises questions about the quality of training, incentives to publish, and even the integrity of the researchers. But in the face of limited resources, researchers might simply be trading lower cost for lower precision. Moreover, this trade off might be optimal given everything the researcher knows. This paper shows that resource constraints should be part of the discussion for how to improve behavioral science.

Limitations and Conclusion

The data provide a unique window onto the way researchers work. Still, the data describe experiments at only one university behavioral lab. The issues raised by this study are relevant to other institutions, but it is unclear how to generalize specific findings to other labs. Cost sensitivity should vary across institutions due to differences in researchers and resources. The data in this study come from the lab's participant scheduling system, and are thus typical of what other labs might have available. It should be straightforward to repeat this analysis at other labs.

Another limit on these results is that the evidence relating higher costs to smaller samples is indirect. The data are observational, and the analysis cannot rule out all alternative explanations for every effect. Cost sensitivity provides a straightforward explanation for most patterns seen in the data. But in some cases the explanation isn't as straightforward. Future work should try to study researcher behavior in more controlled settings.

A third limit is that the archival data only describe studies conducted onsite at the lab. Studies by the same researchers using online panels or conducted in the field are not part of the data. The estimates for price sensitivity thus pertain to lab use only. The same researchers in different settings might exhibit different degrees of cost sensitivity. In particular, reimbursement at this lab is higher and data collection slower than when using MTurk.

A final caveat to these results is that the data do not tell us about studies that didn't make it to the lab. They lack the information we would need to understand when and why researchers run some studies but not others. Thus, the estimated price elasticity pertains only to studies important

enough to actually get run. The studies that do get run are typically those that are of the highest value to researchers. Thus the estimated price elasticity is probably a lower bound (in magnitude) compared to an estimate that were to somehow include the decision of whether to run a study or not.

In spite of these limits, these results show the importance of understanding how costs affect science. The idea that researchers are consumers of information leads to new insights about how they work. By applying theories and tools from consumer research, we can learn even more. This might lead not only to a better understanding of researcher, but also to new ideas for how to improve science. Pushing this metaphor even further may yield new solutions to old problems.

Sample Selection Procedure

Prior to deduplicating and joining studies conducted in both the credit and paid pools, the raw data include 547 experiments and 96 researchers in the credit pool's scheduling system, and 351 experiments and 114 researchers in the paid pool's system. The following steps lead to the final sample of 809 experiments and 155 researchers.

1. Researchers registered in both systems are identified based on their full names, usernames, and email addresses. Users for which the minimum Damerau-Levenshtein (DL) distance between any of these values is three or less are considered to be the same researcher registered in both systems.[4] There are 55 such users, resulting in 155 unique researchers.

2. Starting with 898 unique studies registered in the two systems, 53 had total sample sizes of zero and are removed.

3. To identify which of the remaining 845 studies are duplicated across the two systems, the difference between the first and last dates of data collection, study duration, DL distance between study names, and sets of overlapping researchers associated with each study are considered. This identifies 119 studies registered in both systems as having 1) the same total duration, 2) absolute differences of fewer than five days between either the studies' start or end dates, and 3) either

    (a) Identical study names, but no overlapping researchers (meaning a different researcher registered the study in each of the two systems), or

    (b) Study names with a DL distance less than 40 and at least one researcher associated with both studies.

4. Of the 726 merged experiments, 37 that were not conducted in the lab facilities, such as

[4]The DL distance between two words is analogous to the fewest edits needed to turn one into the other, where an edit could be inserting, replacing, or removing a single character, or swapping the position of two adjacent characters (Damerau, 1964).

those marked as taking place online, at the nearby medical center, or at a local movie theater, are excluded.

5. Registrations for 15 studies are represented by an extra 16 duplicate records corresponding with multi-part studies (all but one are two-part studies, the other is a three-part study). For these studies, the total duration is the sum across all sessions.

6. Among the remaining 673 studies, there are many cases when a study collects data for some period, followed by a longer period of inactivity, and then a resumption of data collection. If at least two weeks of inactivity occurred between observations, it is assumed that researchers re-used an existing registration for what is, in fact, a different study. This procedure yields the 809 studies used for the analysis.

7. Of the 155 unique researchers in step #1 above, 21 are not associated with any of these 809 studies. Their removal yields the 134 researchers used for the analysis.

Student Participation in Credit-Reimbursed Studies by Month

Table 3 shows the correlation in the number of studies each student participated in each month. The negative correlations within academic terms (September–November, January–March, and April–June) arise because students only have to participate in a small number of studies each term (often just one). Thus a student earning all required credits in September has no incentive to participate in October or November. The positive correlations among months at the start or end of each term is the source of selection bias described in the main text.

This behavioral pattern can also be seen via a simple confirmatory cluster analysis. Applying K-means clustering on the same data as above, students are assigned into one of three clusters. For each cluster, the average number of studies participated in each month are reported in Table 4. Cluster C is the largest group and contains most of the student population. Students in cluster C do not participate in many studies (3.3 per student).

Table 3. Correlation in Student Participation in Credit-Reimbursed Studies by Month

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Sep	-0.13	-0.24	0.14	0.33	0.05	-0.10	0.22	-0.06	-0.08
Oct		-0.09	-0.05	0.09	0.20	0.05	0.18	0.15	0.05
Nov			-0.03	-0.06	0.18	0.28	-0.07	0.20	0.21
Dec				0.10	-0.01	0.01	0.07	0.02	0.01
Jan					0.08	-0.02	0.38	0.06	-0.01
Feb						0.18	0.14	0.42	0.18
Mar							-0.03	0.28	0.29
Apr								-0.01	-0.05
May									0.13

Table 4. Average Number of Credit-Reimbursed Studies among Students in each of Three Clusters

Group	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Students
A	1.85	0.95	0.33	0.13	2.19	1.08	0.31	3.27	0.74	0.08	1473
B	0.58	1.04	1.15	0.05	0.61	2.39	1.18	0.60	2.67	0.39	1966
C	0.62	0.52	0.43	0.03	0.31	0.30	0.23	0.37	0.41	0.08	6384

The remaining students engage in considerably more studies (about 11 on average). These students are split roughly equally between clusters A and B. Cluster A includes students who participate in studies at the start of the academic term. Cluster B includes those who participate at the end.

Instrumental Variables Regression

Estimates for the IV models are obtained using two-stage least squares, as implemented in the `AER` R package (Kleiber & Zeileis, 2008). The null hypothesis of weak instruments is unlikely, as R^2 statistics for the first-stage regressions are .27 for use of the credit pool (models 2–3), and .81 (models 2–3) and .82 (model 4) for the amount paid to participants. The largest P-value from first-stage F-tests is $P = .00017$ for c_i in model 3.

The largest Hausman statistic is $P = 0.003$ for model 4, showing evidence against the null that the OLS estimates are consistent (thus IV regression is appropriate). The smallest P-value for the Sargan test is $P = .15$ for model 2, thus failing to show evidence that the instrumental

variables are correlated with the second-stage residuals.

References

Abraham, W. T. & Russell, D. W. (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., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (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., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. doi:10.1002/per.1919

Bakker, M., van Dijk, A., & Wicherts, J. M. (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., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (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., Eliashberg, J., Gatignon, H., & Lodish, L. M. (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

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191.

Fiedler, K., Kutzner, F., & Krueger, J. I. (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. (2016). What has happened down here is the winds have changed. Retrieved from https://web.archive.org/web/20180130080346/http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Gelman, A. & Carlin, J. (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

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

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3), 358–362.

Inman, J. J., Campbell, M. C., Kirmani, A., & Price, L. L. (2018). Our vision for the journal of consumer research: It's all about the consumer. *Journal of Consumer Research*, *44*(5).

International Monetary Fund. (2015). World economic outlook database. Retrieved from https://www.imf.org/external/pubs/ft/weo/2015/01/weodata/index.aspx

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

Kahn, B. E. (2007). Moving the needle: Can ACR help increase our research productivity. *Advances in Consumer Research*, *34*, 1.

Kleiber, C. & Zeileis, A. (2008). *Applied econometrics with R*. ISBN 978-0-387-77316-2. New York: Springer-Verlag. Retrieved from https://CRAN.R-project.org/package=AER

Kraemer, H. C., Gardner, C., Brooks III, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, *3*(1), 23.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (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

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

Meyvis, T. & Van Osselaer, S. M. (2017). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research*, *44*(5), 1157–1173.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. doi:10.1126/science.1245317

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

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of consumer research*, *28*(3), 450–461.

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. & Ball, A. D. (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

Sedlmeier, P. & Gigerenzer, G. (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., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (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, April 4). Mturk vs. the lab: Either way we need big samples. Retrieved from http://web.archive.org/web/20170313102117/http://datacolada.org/18

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (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., Nelson, L. D., & Simonsohn, U. (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

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*, *54*(285), 30–34.

Winkens, B., Schouten, H. J., van Breukelen, G. J., & Berger, M. P. (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