Using Field Experiments in Accounting and Finance

Eric Floyd and John A. List

Rice University, University of Chicago and NBER

The gold standard in the sciences is uncovering causal relationships. A growing literature in economics utilizes field experiments as a methodology to establish causality between variables. Taking lessons from the economics literature, this study provides an "A-to-Z" description of how to conduct field experiments in accounting and finance. We begin by providing a user's guide into what a field experiment is, what behavioral parameters field experiments identify, and how to efficiently generate and analyze experimental data. We then provide a discussion of extant field experiments that touch on important issues in accounting and finance, and we also review areas that have ample opportunities for future field experimental explorations. We conclude that the time is ripe for field experimentation to deepen our understanding of important issues in accounting and finance.

Keywords: Field experiments, causality, identification, experimental design, replication
JEL codes: C00, C9, C93, G00, M4, M5

We would like to thank Douglas Skinner, Brian Akins, Hans Christensen, Yael Hochberg, Michael Minnis, Patricia Naranjo, Brian Rountree, Haresh Sapra, and participants at the JAR 50th Annual Conference for helpful comments and suggestions. We appreciate excellent research assistance from Seung Lee, Rachel Yuqi Li, Ethan Smith, and Rustam Zufarov.

1. Introduction

Rolling weighted balls down a shallowly inclined ramp, Galileo used scientific experiments to prove that matter moves vertically at a constant rate (regardless of mass) due to gravitational effects. Ever since, the experimental approach has been a cornerstone of the scientific method. Whether it was Sir Isaac Newton conducting glass prism experiments to educate himself about the color spectrum or Charles Darwin and his son Francis using oat seedlings to explore the stimuli for phototropism, researchers have rapidly made discoveries since Galileo laid the seminal groundwork. Scientists have even taken the experimental method from the lab to the field. In one classic 1882 example, Louis Pasteur designated half of a group of 50 sheep as controls and treated the other half using vaccination. All animals then received a lethal dose of anthrax. Two days after inoculation, every one of the 25 control sheep was dead, whereas the 25 vaccinated sheep were alive and well. Pasteur had effectively made his point!

Increasingly, social scientists have turned to the experimental model of the physical sciences as a method to understand human behavior. Much of this research takes the form of laboratory experiments in which volunteers enter a research lab to make decisions in a controlled environment (see Bloomfield, Nelson, and Soltes [2015] in this special issue). Over the past two decades, economists have increasingly left the ivory tower and made use of field experiments to explore economic phenomena, studying actors from the farm to the factory to the board room (see Harrison and List [2004]). Much different from experimentation in the hard sciences, field experimenters in economics typically use randomization to estimate treatment effects. And unlike laboratory experiments in the social sciences, field experiments are typically conducted in naturally occurring settings, in certain cases extracting data from people who might not be aware that they are experimental participants.

In this way, field experiments provide a useful bridge between laboratory and naturally occurring data in that they represent a mixture of control and realism usually not achieved in the lab or with uncontrolled data. This unique combination provides the researcher with an opportunity to address questions that heretofore have been difficult to answer. Importantly, after grasping the interrelationships of factors in the chosen field setting, the field experimenter then uses theory to understand more distant phenomena that have the same underlying structure. When this is achieved, the deep rewards of field experimentation are realized.

In this study, we leverage what we have learned in the economics literature to provide a "how-to guide" for developing, implementing, and executing efficient and robust field experiments in accounting and finance. Interestingly, accounting research has traditionally focused on the measurement and auditing of firm performance information that is for use by both internal management for decision making and external users of financial information, such as analysts and investors. Within the field of accounting research, inquiries have ranged from the examination of accounting and audit quality on a firm's cost of capital to understanding the effects of executive compensation contract provisions on CEO incentives. Because of the institutional nature of accounting, research in these areas has typically been archival, focusing on the use of financial market databases, such as *Compustat*, to provide empirical evidence. Yet despite the benefits of using empirical settings that encompass a majority of the institutions of potential interest, considerable challenges remain with respect to treatment identification. Large-scale archival accounting research is often plagued by the absence of exogenous variation, thereby limiting the degree to which researchers can effectively demonstrate causality (Gow, Larcker, and Reiss [2015]).

At the other end of the spectrum, experiments in accounting have tried to address these issues in the laboratory. Lab experiments have extended across multiple areas of accounting research (Libby, Bloomfield, and Nelson [2002]) and have identified several interesting phenomena that are difficult to document empirically in archival studies. Beyond providing a playbook for conducting a field experiment, this article bridges the gap between archival and lab work by showing the promise of field experiments in accounting and finance, which in many cases permit the researcher to establish a tighter link between theory and empirics.

In the next section, we begin by summarizing the various empirical approaches one can use when trying to establish causality. We then define the various field experiments, provide a description of the behavioral parameters that they estimate, and summarize how to implement a field experiment—from theoretical construction to executing the optimal experimental design. From there, we discuss three interconnected issues related to the building of knowledge from field experiments. The issues revolve around appropriate hypothesis testing, how to update one's priors after conducting a field experiment, and the role of replication. We conclude with a discussion of extant field experiments that touch on important accounting and finance issues, and we review areas that are ripe for future field experimental explorations.

2. Empirical Approaches

The empirical gold standard in the sciences is to identify a causal effect of some variable (or set of variables) on another variable. For example, measuring the effect of a new governmental law on the reporting of corporate information or how various compensation schemes affect employee productivity are queries for the scientist interested in causal relationships. The difficulty that arises in establishing causality is that either the treatment is given or it is not—we never directly observe what would have happened in the alternative state. This problem, one of generating the appropriate counterfactual, combined with the fact that in the real world there are deep market complexities and simultaneously many moving parts, has led scholars interested in empiricism to focus on the analysis of naturally occurring data. This is where we begin our discussion, which draws heavily from the related work of Harrison and List [2004], List [2006], and Al-Ubaydli and List [2013]. To do so, we follow their analyses closely and reproduce some of the figures used in their discussions to make our points.

Constructing the proper counterfactual is the key behind any evaluation method. Without loss of generality, define y_1 as the outcome with treatment and y_0 as the outcome without treatment, and let $T = 1$ when treated and $T = 0$ when not treated. The treatment effect for person i can then be measured as $\tau_i = y_{i1} - y_{i0}$. The major problem, however, is one of a missing counterfactual—person i is not observed in both treated and non-treated states.

To estimate the missing counterfactual, social scientists typically follow an age-old process of research: they have an idea, race back to their office to write down a model, download mounds and mounds of data (increasingly from the Internet), and proceed to estimate regression models (i.e., beat up the data until they "talk"). In this spirit, scholars develop a deductive model "assuming rationality" and then run a regression to "test" that deductive model in a way that satisfies almost no one. One reason for the lack of satisfaction is the range of assumptions that must be made in the empirical analysis to establish causality. In many cases, this amounts to asking the following question: what assumptions need to be made to assume that these naturally occurring data arose from a valid experiment?

We provide Figure 1 to place this approach in perspective. The easternmost portion of Figure 1 highlights a handful of popular empirical approaches using naturally occurring data. Consider natural experiments (difference-in-differences models), a workhorse in the accounting literature. Identification in natural experiments results from a difference-in-differences regression model: $y_{it} = X_{it}\,\beta + T_{it}\,\tau + \eta_{it}$, where i indexes the unit of observation, t indexes units of time (e.g., years), y_{it} is the outcome of interest, X_{it} is a vector of controls, T_{it} is a binary treatment variable, $\eta_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$, and τ is measured by comparing the difference in outcomes before and after for the treated group with the before and after outcomes for the non-treated group.[1] A major identifying assumption in this case is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with T_{it}, and that selection into treatment is independent of the temporary individual-specific effect.

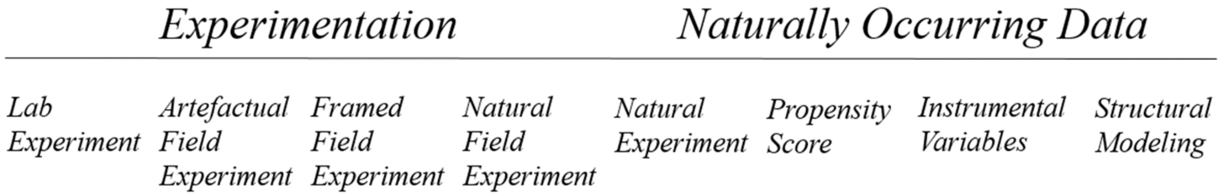

Figure 1: The "Field Experiment Bridge"

One recent example in this spirit is Christensen, Hail, and Leuz [2015]. Their paper uses the staggered entry-into-force dates of two securities regulation directives in the European Union. They use the dates that these regulations were implemented to provide evidence for the causal relationship between regulation and capital market consequences (in this case, market liquidity). In comparison to papers that use a similar methodology in accounting research with limited effectiveness, one of the strengths of the Christensen, Hail, and Leuz [2015] study is the extent that the authors painstakingly address the identifying assumptions described above.

[1] Note that in this formulation, the analyst is assuming a common treatment effect, τ, rather than a heterogeneous treatment effect, τ_i. Use of a random-coefficients econometric model in this framework permits estimation of heterogeneous treatment effects.

Useful alternatives to the difference-in-differences approach include the method of propensity score matching (PSM) developed by Rosenbaum and Rubin [1983]. Again, if both states of the world were observable, the average treatment effect, τ, would equal $\overline{y}_1$-$\overline{y}_0$. However, given that only y_1 or y_0 is observed for each observation, unless assignment into the treatment group is random, generally $\tau \neq \overline{y}_1 - \overline{y}_0$.

The solution that Rosenbaum and Rubin [1983] advocate for is to find a vector of covariates, Z, such that $y_1, y_0 \perp T | Z$, $\text{pr}(T=1|Z) \in (0,1)$, where ⊥ denotes independence. This assumption is called the "conditional independence assumption," and it intuitively means that given Z, the non-treated outcomes are what the treated outcomes would have been had they not been treated. Or, equivalently, that selection into treatment occurs *only* on observables. If this condition holds, then treatment assignment is said to be "strongly ignorable" (Rosenbaum and Rubin [1983], p. 43). To estimate the average treatment effect (on the treated), only the weaker condition $\text{E}(y_0|T=1, Z) = \text{E}(y_0|T=0, Z) = \text{E}(y_0| Z)$, $\text{pr}(T=1|Z) \in (0,1)$ is required. Thus, the treatment effect is given by $\tau = \text{E}(\overline{y}_1 - \overline{y}_0|Z)$, implying that conditional on Z, assignment to the treatment group mimics a randomized experiment.[2]

Christensen, Hail, and Leuz [2015] use PSM to provide a robustness test for their difference-in-differences results. Their empirical design also utilizes a within-country analysis that allows them to compare the effects of regulation between a treated group and a non-treated group

[2] Several aspects of the approach are "left on the sidelines" in this necessarily brief discussion. For example, for these conditions to hold the appropriate conditioning set, Z, should be multi-dimensional. Second, upon estimation of the propensity score, a matching algorithm must be defined in order to estimate the missing counterfactual, y_0, for each treated observation. The average treatment effect on the treated (TT) is given by $\tau_{TT} = \text{E}(\text{E}(y_1 \,|\, T=1, \text{p}(Z)) – \text{E}(y_0 \,|\, T=0, p(Z))) = \text{E}(\text{E}(y_1 - y_0 \,|\, \text{p}(Z)))$, where the outer expectation is over the distribution of $Z \,|\, T = 1$. These and other issues are discussed in Al-Ubaydli and List [2013]. We return to some of these issues below when discussing generalizability.

within each country. Importantly, this differs from their empirical design described above in which treatment effects are measured between-country (i.e., non-treated countries are used as controls). In the within-country design, the authors utilize the fact that certain markets within a country are unregulated, which allows comparison to regulated markets affected by the securities regulation. Unsatisfied with the assumption that these two markets are equivalent on all dimensions except for treatment, the authors utilize PSM to condition on observables (e.g., total assets and return on assets). The strength of these inferences relies on the assumption that the conditional independence assumption is satisfied.

Other popular methods of measurement using naturally occurring data include the use of instrumental variables (IV; see Rosenzweig and Wolpin [2000]) and structural modeling. Assumptions of these approaches are well documented and are not discussed further here (see Blundell and Costa Dias [2002] for a useful review). One recent example of an interesting IV study from the accounting literature is Minnis [2011]. Furthermore, an interesting structural approach in accounting can be found in Zakolyukina [2015]. One benefit of estimating a structural model is that policy counterfactuals are readily available. Below, we discuss recent papers in economics that leverage field experiments to estimate structural models.

At the other end of the empirical spectrum in Figure 1 are laboratory experiments. Current practice concerning the design of a controlled laboratory experiment in economics largely relies on Smith [1980, 1982] and Wilde [1980], as does the following brief summary (see also List [2006]). The lab experimenter's goal is to create a small-scale environment in the laboratory where adequate control is maintained. Such control is necessary to ensure appropriate measurement of treatment effects. An economic environment consists of a set of agents (1,….*n*) and commodities (1,…*k*). Each agent is described by a utility function, u_i, a technology, or knowledge, endowment,

K_i, and a commodity endowment, w_i. Each agent is therefore described by $\varepsilon_i(u_i,K_i,w_i)$, and the microeconomic environment is defined by the collection of agents, $\varepsilon = (\varepsilon_i\ldots\ldots\varepsilon_n)$.

To complete the microeconomic environment, the experimenter specifies the institutional setting, I, which includes the appropriate message space, M; the allocation rules, H; and other relevant characteristics of the specific institution of interest. Thus, the experimental system, $S = (\varepsilon, I)$, is composed of the microeconomic environment and the institution. Agents, who are assumed to possess consistent preferences and to make decisions that maximize their own well-being, choose messages, and the institution determines allocations via the governing rules.

In order to reliably measure the behavioral principles among preferences, institutions, and outcomes, experimenters have proposed a set of sufficient conditions for a valid controlled microeconomic experiment (Smith [1982], Wilde [1980]). These five "precepts" are now commonly used to motivate good experimental practice. They are as follows: nonsatiation (more money is preferred to less), saliency (actions are linked to rewards), dominance (reward structure dominates subjective costs), privacy (each subject is given his/her own payoff structure), and parallelism (experimental environment mirrors the environment of ultimate interest).

From this description, the power of an ideal laboratory experiment readily becomes apparent. In some sense, laboratory experimentation is the most convincing method of creating the counterfactual because it directly constructs a control group via randomization. Such randomization acts as an instrumental variable and therefore allows the analyst to make strong causal statements within the domain of study. We point the interested reader to Bloomfield, Nelson, and Soltes [2015] in this volume (also see Libby, Bloomfield, and Nelson [2002]), which provides an excellent discussion of lab experiments in accounting. Between laboratory experiments and methods using naturally occurring data in Figure 1 are the various types of field experiments, which we turn to next.

3. What Is a Field Experiment?

Field experiments represent a movement to take the data-generation process beyond the walls of the laboratory. Two decades ago, the primary data generators in economics were lab experimentalists. The past 20 years has witnessed an explosion of creative ways to generate data in the field. Harrison and List [2004] propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate. Using these factors, they discuss a classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the lab to the field.

According to this classification scheme, the most minor departure from the typical laboratory experiment is the "artefactual" field experiment (AFE), which mimics a lab experiment, except that it uses "non-standard" subjects—see Figure 1. Such subjects are "non-standard" in the sense that they are not college students but participants drawn from the market of interest. This type of experiment represents a useful type of exploration beyond traditional laboratory studies. AFEs have been fruitfully used in many economic applications, including in public economics, environmental economics, labor economics, and industrial organization, as well as to explore issues in accounting and finance.

Within accounting and finance, Libby and Brown [2013] use an AFE to determine whether auditors have less tolerance for misstatements when the associated income statement numbers are reported at a more disaggregated level. The experiment is best classified as an AFE, as opposed to a lab experiment, because it utilizes experienced US auditors as the subject pool.[3] This is crucial because the research question is directly interested in how professional auditors behave, and student subjects might not provide a viable sample, given the questions addressed. One view of the current experimental literature in accounting is that the most significant move toward field experimentation has been in the use of AFEs.

Moving closer to how naturally occurring data are generated, Harrison and List [2004] denote a framed field experiment (FFE) as the same as an AFE but with field context in the commodity, task, stakes, or information set that the subjects can use—see Figure 1. This type of experiment is important in the sense that myriad factors might influence behavior, and by progressing slowly toward the environment of ultimate interest, one can learn about whether—and to what extent—such factors influence behavior in isolation.

FFEs represent a very active type of field experiment in the past few decades. Alevy, Haigh, and List [2007] provide an example of a framed field experiment used in financial markets. The authors show that financial market professionals behave in very different ways than students in an information cascade experiment. This finding provides important context for prior descriptions of using professional subjects in accounting research. These descriptions suggest that researchers should avoid using professional subjects and settings unless they are necessary to answer the research question being investigated (Libby, Bloomfield, and Nelson [2002]).

[3] One could also make an argument to classify this as a framed field experiment (more on these next) because the auditors are performing tasks that are likely similar to what they typically undertake. This serves to illustrate that different types of field experiments are not perfectly distinct.

These descriptions are correct in noting that using professionals from the field is unnecessary (and more costly) if the theory being tested is independent of the characteristics of the field or if individual experience and selection is believed to be unimportant (see List [2003], [2004], and [2011] as examples where market experience and selection are very important). However, Alevy, Haigh, and List [2007] suggest that within financial markets, the deviation of subject behavior between the lab and field could be of particular concern. This could be of even further consideration to accounting research, which is particularly sensitive to institutional details (e.g., auditing). Importantly, this should not be interpreted as a reason *not* to conduct lab experiments. Instead, the proper interpretation of these studies is that framed field experiments provide an additional source of information that is not typically found within the lab. As such, a comparison of results across lab, artefactual, and framed field experiments allows us to understand whether representativeness of the population and representativeness of the situation are important.

Finally, a natural field experiment (NFE) is the same as an FFE in that it occurs in the environment where the subjects naturally undertake these tasks, but in an NFE, the subjects *do not know* that they are participants in an experiment—see Figure 1.[4] Such an exercise is important because it represents an approach that combines the most attractive elements of the experimental method and naturally occurring data: randomization and realism. In addition, it tackles a selection problem that is often not discussed concerning the other types of experiments, as discussed below.

[4] This raises the issue of informed consent. For a discussion on this and related issues, see List [2008] and Levitt and List [2009].

NFEs have recently been used to answer a wide range of questions in economics, including topics as varied as measuring preferences (List [2003]) and how one can manage an on-line shopping experience (Hossain and Morgan [2006]). Recently, Hallsworth, List, and Metcalfe [2014] use large-scale natural field experiments to examine the factors that influence timely payment of taxes. In partnership with the United Kingdom tax collection authority (Her Majesty's Revenue and Customs (HMRC)) and the UK Cabinet Office's Behavioural Insights Team (BIT), the authors randomize five messages across roughly 100,000 individual taxpayers: three norm-based messages and two public services messages. A control group receives a standard letter with no persuasive message of this kind. Empirical results show that social norm and public services messages increase the likelihood of individuals paying their declared tax liabilities. The most successful message produces a treatment effect of 5.1%; this improvement is sustained until at least 70 days after the letters are issued. The authors estimate that more than £3 million is accelerated during the 23-day sample period due to the messages in the first field experiment.

Of course, the taxonomy in Figure 1 leaves gaps, and certain studies may not fall neatly into such a uni-dimensional classification scheme, as we discussed with Libby and Brown [2013]. Yet such an organization highlights what is necessary in terms of scientific discovery to link controlled experimentation in the lab to naturally occurring data. Importantly, they also provide different behavioral parameter estimates, which we turn to next.

4. What Behavioral Parameters Do Lab and Field Experiments Estimate?

Following our earlier notation, define y_1 as the outcome with treatment, y_0 as the outcome without treatment and let $T=1$ when treated and $T=0$ when not treated. Also, assume that $p=1$ indicates participation in the experiment, while $p=0$ indicates non-participation. That is, people who agree to enroll in the experiment have $p=1$, and others have $p=0$. If one is interested in the mean differences in outcomes, then the treatment effect that can be captured is given by a simple comparison of the *p=1* group:

$$t = \mathrm{E}(\tau|p = 1) = \mathrm{E}(y_1 - y_0|p = 1)$$

Yet in our experience in the field, agencies, firms, and laypeople who discuss results from evaluations of programs typically report the following treatment effect:

$$t' = \mathrm{E}(y_1|\, p = 1) - \mathrm{E}(y_0|\, p = 0)$$

This reported effect represents a potentially misleading measure because it is comparing the mean outcome for two potentially quite different populations. To see the difference between t and t', simply add and subtract $\mathrm{E}(y_0|\, p{=}1)$ from *t'*, yielding the following:

$$t' = \underbrace{\mathrm{E}(\tau|p = 1) = \mathrm{E}(y_1 - y_0|p = 1)}_{t} + \underbrace{\mathrm{E}(y_0|\, p = 1) - \mathrm{E}(y_0|\, p = 0)}_{\delta}$$

where δ is the traditional selection bias term. This bias is equal to the difference in outcomes across the non-participants (*p=0*) and participants (*p=1*) in the *non-treated state*.

This equation is illustrative because it shows clearly how selection bias, as is typically discussed in the literature, relates to outcomes in the non-treated state. Consider a social program for underprivileged children. You might ask, which parents are most likely to sign up their children for the program? One logical answer is that it is those parents who care most deeply about their children's lifetime outcomes. If this is correct, then their children might have better outcomes in the non-treated state. In this case, such a selection bias causes the second term to be greater than zero because $\mathrm{E}(y_0|\, p{=}1) > \mathrm{E}(y_0|\, p{=}0)$, leading the program to report a treatment effect that is too optimistic (biased upward). In our travels, we have found that this problem—one of not constructing the proper control group—is ubiquitous.

To avoid this sort of selection bias, it is necessary for the randomization and identification of the treatment effect to occur over the *p=1* group, yielding a treatment effect estimate of the mean outcome differences between treated and non-treated for the *p=1* group. Letting *D=1(0)* denote those randomized into treatment (non-treatment):

$$t = \mathrm{E}(y_1 | D = 1\ AND\ p = 1) - \mathrm{E}(y_0 |\ D = 0\ AND\ p = 1)$$

At this point, it is instructive to pause and ask how to interpret the meaning of this treatment effect. In short, this is the treatment effect that laboratory experiments, as well as AFEs and FFEs, estimate. Given that randomization was done appropriately, this is a valid treatment effect estimate for the *p=1* population.

For this effect to generalize to the *p=0* population, however, further assumptions must be invoked. Most importantly, the effect of treatment cannot differ across the *p=1* and *p=0* groups. If, for instance, a person has a unique trait that is correlated with participation status and correlated with the outcome variable, such a generalization is frustrated. In our program example above, it might be the case that parents who believe that the program will have a positive effect on their child are more likely to enroll. If this is indeed true, then it would not be appropriate to generalize the effect from the *p=1* group to the *p=0* group. By doing so, the analyst would be reporting an effect of treatment for the *p=0* group that is biased upward.

This effect—which Al-Ubaydli and List [2013] denote as Treatment Specific Selection Bias—is quite distinct from the traditional selection bias discussed in the literature and shown above. Whereas the standard selection bias relates to outcomes of the *p=1* and *p=0* groups in the non-treated state, Treatment Specific Selection Bias relates to outcomes of the *p=1* and *p=0* groups in the *treated* state.

So how can the researcher avoid such a bias? NFEs represent an excellent option. Because in NFEs subjects are unaware that they are taking part in an experiment, NFEs naturally resolve any bias issues. In this case, there is no *p=1* or *p=0* group: subjects are randomly placed into treatment or control groups without even knowing. Thus, NFEs exclude both the typical selection problem discussed in the literature and preclude Treatment Specific Selection Bias. Indeed, it also rids us of other biases, such as randomization bias and any behavioral effects of people knowing that they are taking part in an experiment (see Levitt and List [2007]).

The very nature of how the NFE parameter is estimated reveals the mistake that many people make when claiming that the laboratory environment offers more “control” than a field experiment. There are unobservables in each environment, and to conclude *ex ante* that certain unobservables (field) are more detrimental than others (lab) is missing the point. This is because randomization balances the unobservables—whether myriad or a few. Thus, even if one wished to argue that background complexities are more severe in one environment than the other, there really is little meaning—one unobservable can do as much harm as multiple unobservables. The beauty behind randomization is that it handles the unobservability problem, permitting a crisp estimate of the causal effect of interest.

Furthermore, Al-Ubaydli and List [2015] show that NFEs actually provide *more* control than lab experiments, AFEs, and FFEs concerning the selection of subjects. Their study shows that—contrary to conventional wisdom—NFEs can offer experimenters *more* control than other experiment types, specifically in terms of the participation decision. They also suggest that in certain cases, relying on laboratory experiments can create biases in our estimates of causal effects due to heterogeneity among subjects; when researchers estimate causal effects using samples that differ (observably or unobservably) from the target population, they risk an extrapolation bias as a result of potential heterogeneity in causal effects (Heckman [2005]). The only definitive remedy is to consider redesigning the experiment in a manner that ensures experimental participation by members of the target group.

Accordingly, a more accurate way of describing the difference in control between laboratory and NFEs is as follows: laboratory experiments afford researchers greater control *over the physical environment and the nature of permissible interactions (using induced values, for example—see List [2004])*, but they offer researchers less control *over the nature of the experimental participants*.

5. I Have Decided to Run a Field Experiment—Now What?

Given the potential appeal of field experiments described above, of practical importance is how one goes about conducting a field experiment. In accounting and finance research, scholars are typically interested in addressing questions that are either within or across firms (or managers of firms). Consequently, designing experiments that incorporate firms is a critical component in the potential use of field experiments in accounting and finance. Naturally, this issue creates additional hurdles in being able to implement an effective field experiment.

List [2011] outlines the following 14 important tips of practical importance for conducting field experiments with firms:

1. Use economic theory to guide your design and as a lens to interpret your findings.
2. Be an expert about the market that you are studying.
3. Have a proper control group.
4. Obtain correct sample sizes.

5. Have a champion within the organization—the higher up, the better.

6. Understand organizational dynamics.

7. Organizations that have "skin in the game" are more likely to execute your design and use your results to further organizational objectives.

8. Run the field experiment yesterday, rather than tomorrow.

9. Change the nature and discussion of the cost of the experiment.

10. Make clear that you do not have all the answers.

11. Be open to running experiments that might not pay off in the short run.

12. Don't be captured by the organization.

13. Understand fairness concerns.

14. Always obtain Institutional Review Board (IRB) approval.

Although we do not reproduce a discussion of each tip here, several important points are worth mentioning.

The first relates to the use of theory in experimental design and data interpretation—#1 on List's 14 tips. One should always keep in mind that theory is portable; empirical results in isolation offer only limited information about what is likely to happen in a new setting—be it a different physical environment or time period. Together, however, theory and experimental results provide a powerful guide to situations heretofore unexplored. In this way, experimental results are most generalizable when they are built on tests of economic theory.

Consider a recent example in the economics literature that revolves around understanding why people give to charitable causes (DellaVigna, List, and Malmendier [2012]). The authors begin with a structural model that admits two broad classes of motivations for giving: altruism (including warm glow) and social pressure to give. The two motivations have very different welfare implications. The model dictates the set of experimental treatments necessary to parse the underlying motivation for giving. Using the theory as a guide, the authors use a door-to-door fundraising drive, approaching more than 7,000 households to test their theory.

Combining their structural theory and the data drawn from their NFE, the authors quantitatively evaluate the welfare effects for the giver and decompose the share of giving that is due to altruism versus social pressure. In this way, the empirics and theory are intertwined in a manner that is rare in the literature but ultimately of great importance for testing theory and making policy prescriptions. The authors report that both altruism and social pressure are important determinants of giving in this setting, with stronger evidence for the role of social pressure. As a result of having a structural model, the authors can provide welfare policy counterfactuals and report that half of donors derive negative utility from the fundraising interaction and would have preferred to sort out of the interaction. We view this as an important future of field experiments, particularly as a means to provide exogenous variation to estimate structural models (see DellaVigna et al. [2015] on why people vote in elections as an additional example).

Related to having a theory that guides your experiment is #2 in List's list: have a deep understanding of the market of interest. This is perhaps the most important insight that we have gained from more than 20 years of running field experiments. As a sports card dealer running NFEs in the early 1990s, List needed to understand the inner workings of the market, in the sense that he had detailed knowledge of the underlying motivations of the actors in the market—buyers, sellers, third-party certifiers, and organizers. This was quite beneficial in crafting designs in which the incentives would be understood and interpreted correctly, as well as in generating alternative hypotheses and understanding how to interpret the experimental data in light of the theoretical considerations. In sum, this understanding is necessary to go beyond AB testing (A *causes* B) and provide the "whys" underlying observed data patterns.

When partnering with firms, organizational dynamics are an extremely important consideration. Without an understanding of a firm's dynamics, it is much easier for members of the organization to dismiss the credibility of the experiment. Even with this information in hand, researchers must be aware of the incentives of the organization. First, having an influential member of the organization support the experiment is invaluable. Although this is not surprising, it poses a less obvious question: how does one convince an executive of a company to agree to run a field experiment?

The natural response is for the researcher to be thorough about the potential outcomes of the experiment. As one would expect, organizations are much more willing to partner with a research team if the researcher can present a justifiable case that the results of the experiment will ultimately provide some benefit to the firm. Put differently, an organization is less likely to help researchers with the goal of testing nuanced theory than they are researchers who present a clear idea of how the experiment will lead to increased firm profits or enhanced customer or worker experiences.

In short, though field experiments provide unique opportunities to develop economic insight, they also come with their own practical considerations. The more the researcher is aware of these considerations, the greater the researcher's chance of being able to successfully execute a field experiment. We strongly urge the reader interested in conducting her own field experiments to consult List [2011] for a more complete discussion of these issues.

Finally, one prominent reason that field experiments fail is because they were ill powered from the beginning (Tip #4). This stems from the fact that experimentalists do not pay enough attention to the power of the experimental design—whether it be that clustering was not accounted for or other potential nuances were ignored. We turn to this issue now.

6. Nuts and Bolts of Design

Actually designing the data generation process is an important, oft-neglected aspect of field experimentation.[5] This is especially true when one considers that modification is extremely difficult once the experiment has begun. This contrasts with lab experiments, in which it is often feasible for experimenters to replicate an experiment multiple times. This is the reason for optimal sample size considerations being placed #4 on List's list. Scholars have produced a variety of rules of thumb to aid in experimental design. List, Sadoff, and Wagner [2011] and List and Rasul [2011] summarize some of these rules of thumb for optimal design, and we follow those discussions closely here.

To provide a framework to think through optimal sample size intuition, we continue with the notation introduced above where there is a single treatment T that results in outcomes y_{i0} and where $y_{i0}|X_i{\sim}\mathrm{N}\left(\mu_0,\sigma_0^2\right)$ and y_{i1}, where $y_{i1}|X_i{\sim}\mathrm{N}\left(\mu_1,\sigma_1^2\right)$. Because the experiment has not yet been conducted, the experimenter must form beliefs about the variances of outcomes across the treatment and control groups, which may, for example, come from theory, prior empirical evidence, or a pilot experiment. Beyond ensuring that the protocol is manageable and

[5] We do not discuss effective treatment design in this discussion for brevity. Testing several interventions, as opposed to one intervention, induces a tradeoff for the researcher. On the one hand, it helps the researcher to identify the mechanism in which the treatment operates. On the other hand, it potentially limits the power of the experiment to identify the primary effect. This issue is discussed further in the auditing section later in the paper.

understandable for the subjects, the pilot experiment provides information on how treatments affect behavior, which is a key input into the experimental design.

In this way, the pilot experiment provides information for the experimenter that is used to determine the minimum detectable difference between mean control and treatment outcomes, μ_1-μ_0=δ, that the experiment can detect. In this notation, δ is the minimum average treatment effect, $\bar{\tau}$, that the experiment will be able to detect at a given significance level and power. Finally, we assume that the significance of the treatment effect will be determined using a parametric t-test.

The first step in calculating optimal sample sizes requires specifying a null hypothesis and a specific alternative hypothesis. Typically, the null hypothesis is that there is no treatment effect—i.e., that the effect size is zero. The alternative hypothesis is that the effect size takes on a specific value (the minimum detectable effect size). The idea behind the choice of optimal sample sizes in this scenario is that the sample sizes have to be just large enough so that the experimenter (i) does not falsely reject the null hypothesis that the population treatment and control outcomes are equal—i.e., commit a Type I error and (ii) does not falsely accept the null hypothesis when the actual difference is equal to δ—i.e., commit a Type II error.

More formally, if the observations for control and treatment groups are independently drawn and H_0: $\mu_0=\mu_1$ and H_1: $\mu_0\neq\mu_1$, we need the difference in sample means $\bar{y}_1$-$\bar{y}_0$ (which are, of course, not yet observed) to satisfy the following two conditions related to the probabilities of Type I and Type II errors. First, the probability α of committing a Type I error in a two-sided test—i.e., a significance level of α—is given by the following:

$$\frac{\overline{y_1}\text{-}\overline{y_0}}{\sqrt{\frac{\sigma_0^2}{n_o}+\frac{\sigma_1^2}{n_1}}} = t_{\alpha/2} \Rightarrow \overline{y_1}\text{-}\overline{y_0} = t_{\alpha/2}\sqrt{\frac{\sigma_0^2}{n_o}+\frac{\sigma_1^2}{n_1}}$$

where σ_T^2 and n_T for $T = \{0,1\}$ are the conditional variance of the outcome and the sample size of the control and treatment groups, respectively. Second, the probability β of committing a Type II error—i.e., a power of 1-β—in a one-sided test, is given by

$$\frac{\left(\overline{y_1}\text{-}\overline{y_0}\right)-\delta}{\sqrt{\frac{\sigma_0^2}{n_o}+\frac{\sigma_1^2}{n_1}}} = -t_\beta \Rightarrow \overline{y_1}\text{-}\overline{y_0} = \delta - t_\beta\sqrt{\frac{\sigma_0^2}{n_o}+\frac{\sigma_1^2}{n_1}}$$

Using the formula for a Type I error to eliminate $\overline{y_1}\text{-}\overline{y_0}$ from the formula for a Type II error, we obtain

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

It can be easily shown that if σ_0^2=σ_1^2=σ^2—i.e., var(τ_i)$=0$—then the smallest sample sizes that solve this equality satisfy n_0=n_1=n and then

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

If the variances of the outcomes are not equal, this becomes

$$N^* = \left(\frac{t_{\alpha/2}+t_\beta}{\delta}\right)^2\left(\frac{\sigma_0^2}{\pi_o{}^*}+\frac{\sigma_1^2}{\pi_1{}^*}\right),$$

$$\pi_0^* = \frac{\sigma_0}{\sigma_0+\sigma_1}, \pi_1^* = \frac{\sigma_1}{\sigma_0+\sigma_1},$$

where

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

If sample sizes are large enough so that the normal distribution is a good approximation for the t-distribution, then the above equations are a closed-form solution for the optimal sample sizes. If sample sizes are small, then *n* must be solved by using successive approximations.

These equations provide some interesting insights. First, optimal sample sizes increase proportionally with the variance of outcomes, increase non-linearly with the significance level and the power, and decrease proportionally with the square of the minimum detectable effect. Second, the relative distribution of subjects across treatment and control is proportional to the standard deviation of the respective outcomes. This reveals the power of the pilot experiment because if it suggests that the variance of outcomes under treatment and control are fairly similar, there should not be a large loss in efficiency from assigning equal sample sizes to each.

Third, in cases when the outcome variable is dichotomous, under the null hypothesis of no treatment effect, $\mu_0=\mu_1$, one should always allocate subjects equally across treatments. Yet if the null is of the form $\mu_1=k\mu_0$, where $k > 0$, then the sample size arrangement is dictated by *k* in the same manner as in the continuous case. Fourth, if the cost of sampling subjects differs across treatment and control groups, then the ratio of the sample sizes is inversely proportional to the square root of the relative costs. Interestingly, differences in sampling costs have exactly the same effect on relative sample sizes of treatment and control groups as differences in variances.

As List, Sadoff, and Wagner [2011] show, these simple rules of thumb readily fall out of the simple framework summarized above. Yet in those instances where the unit of randomization is different from the unit of observation, special considerations must be paid to correlated outcomes. Specifically, the number of observations required is multiplied by 1+(m-1)ρ, where ρ is the intracluster correlation coefficient and m is the size of each cluster. The optimal size of each cluster increases with the ratio of the within- to between-cluster standard deviation and decreases with the square root of the ratio of the cost of sampling a subject to the fixed cost of sampling from a new cluster. Because the optimal sample size is independent of the available budget, the experimenter should first determine how many subjects to sample in each cluster and then sample from as many clusters as the budget permits (or until the optimal total sample size is achieved). We direct the reader to List, Sadoff, and Wagner [2011] for a more detailed discussion of clustered designs.

A final class of results pertains to designs that include several levels of treatment or, more generally, when the treatment variable itself is continuous, but we assume homogeneous treatment effects. The primary goal of the experimental design in this case is to simply maximize the variance of the treatment variable. For example, if the analyst is interested in estimating the effect of a treatment and has strong priors that the treatment has a linear effect, then the sample should be equally divided on the endpoints of the feasible treatment range, with no intermediate points sampled. Maximizing the variance of the treatment variable under an assumed quadratic, cubic, quartic, etc., relationship produces unambiguous allocation rules, as well: in the quadratic case, for instance, the analyst should place half of the sample equally distributed on the treatment cell endpoints and the other half on the treatment cell midpoint. More generally, optimal design requires that the number of treatment cells used should be equal to the highest polynomial order plus one. Again, we direct the interested reader to List, Sadoff, and Wagner [2011].

After the field experiment has been conducted and the results have been summarized, there is a means by which scientific knowledge accumulates. One fact in the experimental community—whether in economics or psychology—is that there is a shortage of replication experiments. We turn to how that shortage influences how much we can learn from empiricism.

7. Analyzing Data and Building Scientific Knowledge from Field Experiments

We discuss three interconnected issues in this section related to the building of knowledge from field experiments, though the lessons are broadly appropriate for any empirical exercise. The issues revolve around analyzing data from a field experiment after it is conducted (appropriate hypothesis testing), how to update one's priors after conducting a field experiment, and the role of replication of field experimental results in building scientific knowledge.

Multiple-Hypothesis Testing

The approach to analyzing data from experiments is well understood. Most experimenters use both parametric (t-tests and regression analysis) and non-parametric (Wilcoxon signed rank tests, Mann-Whitney test, etc.) approaches, depending on their assumptions about the underlying population. Rather than providing a summary of those basic approaches, which the reader can obtain from any introductory statistics text, we provide a summary of a common shortcoming that we observe in the scientific community: failure to account for multiple-hypothesis testing when doing statistical analyses.

As List, Shaikh, and Xu [2015] discuss, multiple-hypothesis testing refers to any instance in which a family of hypotheses is carried out simultaneously and one has to decide which hypotheses to reject. Within the area of experimental economics, there are three common scenarios that involve multiple-hypothesis testing: i) jointly identifying treatment effects for a set of outcomes (i.e., in education experiments, the scholar is interested in grades, school attendance, and standardized test scores as outcome variables); ii) heterogeneous treatment effects are explored through subgroup analysis (i.e., studies that show gender/age/experience effects or measure effects of geography on behavior); and iii) hypothesis testing is conducted for multiple treatment groups. The third scenario may include two cases: assessing treatment effects for multiple treatment conditions and making all pairwise comparisons across multiple treatment conditions and a control condition.

The intuition behind why it is necessary to adjust for multiple-hypotheses testing is straightforward. In testing any single hypothesis, experimenters typically conduct a t-test where the Type I error rate is set so that we know for each single hypothesis the probability of rejecting the null hypothesis when it is true. When multiple hypotheses are considered together, however, the probability that at least some Type I errors are committed often increases dramatically with the number of hypotheses. Consider a simple illustration for why this is an inferential problem. In the case of one hypothesis test, under standard assumptions, there is a 5% chance of incorrectly rejecting the null hypothesis (if the null hypothesis is true). Yet if the analyst is doing 100 tests at once, where all null hypotheses are true, the expected number of (incorrect) rejections is 5. Assuming independent tests, this means that the probability of at least one incorrect rejection is 99.4%.

How pervasive is the multiple-hypothesis problem in practice? Fink, McConnell, and Vollmer [2014] review all field experiment-based articles published in top academic journals from 2005 to 2010 and report that 76% of the 34 articles that they study involve subgroup analysis, and 29% estimate treatment effects for ten or more subgroups. Furthermore, Anderson [2008] reports that 84% of randomized evaluation papers published from 2004 to 2006 in a set of social science fields jointly test five or more outcomes, and 61% have ten or more outcomes simultaneously tested. Yet only 7% of these papers conduct any multiplicity correction.

So what can be done? List, Shaikh, and Xu [2015] build on work in the statistics literature to present a new testing procedure that applies to any combination of the three common scenarios for multiple-hypothesis testing described above. Under weak assumptions, their testing procedure controls the familywise error rate—the probability of even one false rejection—in finite samples. Their methodology differs from classical multiple testing procedures—such as Bonferroni [1935] and Holm [1979]—in that it incorporates information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. In this way, their procedure is more powerful than previous methods. We urge the reader who does *any* sort of empirical exercise to adjust their standard errors along the lines of List, Shaikh, and Xu [2015] when appropriate (they also provide applicable empirical code).

Updating Priors

As discussed above, most researchers view the key findings from an empirical exercise based solely on observed *p*-values. For instance, the terminology "I can reject the null at the 5% level" is so ubiquitous that it has become part of the standard empiricist's language. To show why this focus on *p*-values is flawed, consider the model in Maniadis, Tufano, and List [2014; MTL hereafter], where they assume that the researcher has a prior on the actual behavioral relationships as follows.[6]

[6] Although such tracks have been covered recently by MTL, we parrot their discussion here because there has been scant mention of this important issue, and it serves to highlight a key virtue of experimentation.

Let n represent the number of associations that are being studied in a specific field. Let π be the fraction of these associations that are actually true.[7] Let α denote the typical significance level in the field (usually α=0.05) and 1-β denote the typical power of the experimental design.[8] As researchers, we are interested in the Post-Study Probability (PSP) that the research finding is true—or more concretely, given the empirical evidence, how sure we are that the research finding is indeed true.

This probability can be found as follows: of the n associations, πn associations will be true, and $(1\text{-}\pi)n$ will be false. Among the true ones, $(1\text{-}\beta)\pi n$ will be declared true relationships, while among the false associations, $\alpha(1\text{-}\pi)n$ will be false positives, or declared true even though they are false. The PSP is simply found by dividing the number of true associations that are declared true by the number of all associations declared true:

[1] $$\mathrm{PSP} = \frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$$

It is natural to ask what factors can affect the PSP. MTL discuss three important factors that potentially affect PSP: (i) how sample sizes affect our confidence in experimental results, (ii) how competition by independent researchers affects PSP, and (iii) how researcher biases affect PSP.

[7] π can also be defined as the prior probability that the alternative hypothesis H_1 is actually true when performing a statistical test of the null hypothesis H_0 (see Wacholder et al. [2004]): that is, π=pr{H_1 is true}.

[8] As List, Sadoff, and Wagner [2011] emphasize, power analysis is not appealing to economists. The reason is that our usual way of thinking is related to the standard regression model. This model considers the probability of observing the coefficient that we observed if the null hypothesis is true. Power analysis explores a different question: if the alternative is true, what is the probability of the estimated coefficient lying outside the confidence interval defined when we tested our null hypothesis?

For our purposes, we can use [1] to determine the reliability of an experimental result. To give an indication of the type of insights that [1] can provide, we plot the PSPs for three levels of experimental power in Figure 2 (from MTL).

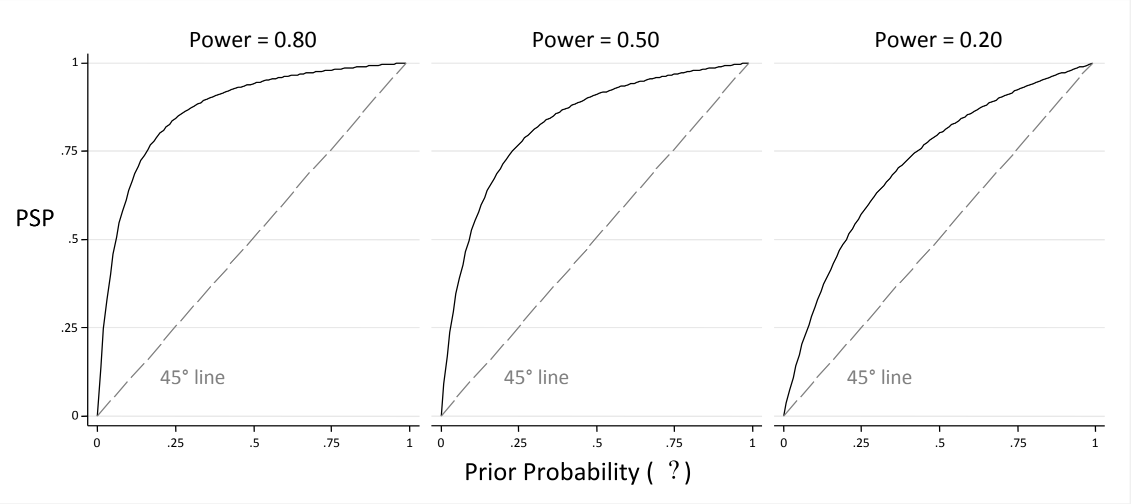

Figure 2: The PSP as a function of power

Upon comparing the leftmost and rightmost panels for the case of π = 0.5, we find that the PSP in the high-power (0.80) case is nearly 20% higher than the PSP in the low-power (0.20) case. This suggests that as policymakers, we would be nearly 20% more certain that research findings from higher-powered exercises are indeed true in comparison to lower-powered exercises.

Importantly, this discussion shows that the common benchmark of simply evaluating *p*-values when determining whether a result is a true association is flawed. The common reliance on statistical significance as the sole criterion for updating our priors can lead to overconfidence about what our results suggest. In this sense, our theoretical model suggests that many surprising new empirical results likely do not recover true associations. The framework highlights that, at least in principle, the decision about whether to call a finding noteworthy, or deserving of great attention, should be based on the estimated probability that the finding represents a true association, which follows directly from not only the observed *p*-value but also the power of the experimental design, the prior probability of the hypothesis, and the tolerance for false positives.

Figure 2 also provides a cautionary view in that it suggests that we should be wary of “surprise” findings (those that arise when π values are low) from experiments with low power because they are likely not correct findings if one considers the low PSP. In these cases, they are likely not even true in the domain of study, much less in an environment that the researcher wishes to generalize.

Replication Is the Key to Building Scientific Knowledge

Because replication is the cornerstone of the experimental method, it is important to briefly discuss the power of replication in this setting, as well. Again, we follow MTL’s [2014] discussion to make our points about how replication aids in making inferences from experimental samples.

As Levitt and List [2009] address, there are at least three levels at which replication can operate. The first and most narrow of these involves taking the actual data generated by an experimental investigation and re-analyzing the data to confirm the original findings. A second notion of replication is to run an experiment that follows a similar protocol to the first experiment to determine whether similar results can be generated using new subjects. The third and most general conception of replication is to test the hypotheses of the original study using a new research design. We focus on the second form of replication using the MTL model described above, yet our fundamental points apply equally to the third replication concept.

Continuing with the notion that the researcher has a prior on the actual behavioral relationships, we follow MTL’s model of replication. MTL’s framework suggests that a little replication can significantly impact the building of scientific knowledge. To illustrate, we consider their Table 5, reproduced here as Table 1. The authors calculate the probability that anywhere from zero to four investigations (the original study and three replications) find a significant result, given that the relationship is true and given that it is false. Then, they derive the PSP in the usual way—as the fraction of the true associations over all associations for each level of replication. The results reported in Table 1 show that with just two independent positive replications, the improvement in PSP is dramatic. Indeed, for studies that report "surprising" results—those that have low π values—the PSP increases more than *threefold* upon a couple of replications.

To understand the values in Table 1, let us consider a simple example. We will assume that we are considering a tobacco control act. The values in the first row of Table 1 can be interpreted as follows: If we believe that there is a 1% chance of the control act having a significant negative effect on smoking behavior before seeing any evidence, upon seeing evidence from a study with power of 0.80, we should update our beliefs to there being a 2% chance of the control act having a negative effect on smoking behavior.

However, if one independent replication also finds a negative effect, we should update our beliefs even more—to 10%. Then, if a second independent replication also finds a negative effect, our beliefs should be updated to 47%. A third independent replication would move our beliefs to 91%. In that scenario, we are now quite confident that there is an effect of the control act.

Table 1—The PSP Estimates as a Function of Prior Probability (*π*), Power, and Number of Replications (*i*)

	Power = 0.80				Power = 0.50				Power = 0.20			
Π	*i=0*	*i=1*	*i=2*	*i=3*	*i=0*	*i=1*	*i=2*	*i=3*	*i=0*	*i=1*	*i=2*	*i=3*
	PSP											
0.01	**0.02**	**0.10**	**0.47**	0.91	**0.02**	**0.10**	**0.45**	0.89	**0.02**	**0.07**	**0.22**	0.54
0.02	**0.05**	**0.19**	0.64	0.95	**0.05**	**0.19**	0.63	0.94	**0.04**	**0.13**	**0.36**	0.71
0.05	**0.12**	**0.38**	0.82	0.98	**0.12**	**0.38**	0.81	0.98	**0.10**	**0.28**	0.60	0.86
0.10	**0.22**	0.56	0.91	0.99	**0.22**	0.56	0.90	0.99	**0.20**	**0.45**	0.76	0.93
0.20	**0.38**	0.74	0.96	1.00	**0.38**	0.74	0.95	1.00	**0.36**	0.64	0.88	0.97
0.35	0.57	0.86	0.98	1.00	0.57	0.86	0.98	1.00	0.55	0.80	0.94	0.98
0.55	0.75	0.93	0.99	1.00	0.75	0.93	0.99	1.00	0.73	0.90	0.97	0.99

As noted above, the third and most general conception of replication is to test the hypotheses of the original study using a new research design. If that exercise is completed, the updating of beliefs is potentially even more powerful than Table 1 suggests (depending on your model of updating). For example, if three studies all find a policy effect and they use different assumptions to formulate the counterfactual, all else equal, that is potentially more informative than three studies that use the same set of assumptions. We should note that we know of no study that examines this speculation.

8. Experiments and Accounting

To provide context for the discussion thus far, we now turn to settings in accounting research that are potentially well suited to the use of field experiments. These settings have all received various levels of attention in the literature, yet each is useful for illustration purposes as they all demonstrate the various benefits and concerns when using field experiments as a methodology to investigate economic theories.

Motivating Example

We begin our discussion in this section by illustrating a particularly salient example of how field experiments can provide novel insights into accounting research. The real effects of accounting disclosure (for a discussion, see Kanodia and Sapra [2015]) are behavioral changes of agents, such as managers, when they anticipate having to disclose information about those behaviors. In the context of accounting, for example, managers might change their investment in firm-level assets when the performance of those assets will subsequently be disclosed to financial markets.

Consider the model in Kanodia, Sapra, and Venugopalan [2004]. The authors analyze corporate investment under two accounting regimes: 1) when intangibles are comingled with operating expenses and 2) when intangibles are disclosed separately. They find that both accounting regimes produce suboptimal (relative to first best) investment. Importantly, they show that disclosing intangible investment separately is only desirable when intangibles are relatively important for firm productivity, and they can be measured precisely. In discussing the potential implications of their findings, the authors state, “We believe that accounting policy on intangibles is better guided by a clear identification of changes in economic decisions, and their consequences, that are caused by a change in the accounting treatment of intangibles.” The natural question is, then, how does one go about identifying these changes in economic decisions?

As is typical in many areas of study, the empirical literature has lagged behind the theoretical literature in this area because of the difficulty in observing real economic decisions, such as investment, for comparable groups of firms in response to an exogenous change in accounting disclosure. As a consequence, many of the theoretical contributions in this area have not been tested and thus remain fertile ground for research.

From an archival researcher's perspective, the task of testing such a theory is extremely challenging. First, very rarely do exogenous disclosure changes occur in an accounting setting while providing for a suitable control group so as to allow for a difference-in-differences approach. Accounting disclosures often apply to the full sample of interest, meaning that a sample for comparison no longer exists. Furthermore, even if a control group does exist, the fact that the experiment is related to disclosure means that information for the control group may not be attainable.

Christensen et al. [2015] is a notable exception. In their study, the authors are able to investigate the effect of mine safety disclosures in financial statements by comparing public firms to private firms. Only SEC-registered firms are required to disclose mine safety information, which provides the authors a potential control group under certain assumptions. Information is obtained for the private group because a government organization (Mine Safety and Health Administration (MSHA)) measures investment and safety outcomes, regardless of whether the firm will disclose information in financial statements. The information that the authors receive from MSHA is also available to the general public via a website. Thus, their experiment is not analyzing the pure effect of mine safety disclosures but instead the incremental effect of mine safety disclosures within financial statements. As a whole, this type of rich data situation is not amenable to many questions that interest accountants.

Other empirical investigations have attempted to analyze the real effects of disclosure but have done so in settings outside of accounting (Leuz and Wysocki [2015]). For example, Jin and Leslie [2003] examine the effects of restaurant hygiene reports on quality. Floyd and Tomar [2015] investigate the effects of grade non-disclosure on student investment in classes and student-employer matching. Both cases illustrate that finding suitable quasi-experiments to investigate the real effects of disclosure can be challenging.

Field experiments provide one avenue in which researchers in accounting may be able to circumvent these challenges. In doing so, field experiments may be able to provide a well-needed bridge between theory and empirics in accounting research. In the Kanodia, Sapra, and Venugopalan [2004] setting, the ideal experiment would randomize accounting rules to different firms and measure investment at the firm level both before and after treatment. This design would allow the researcher to establish causality between disclosure requirements and changes in real activities, such as investment in intangibles. Furthermore, researchers could explore if the observed treatment effect varies with the relative proportions of intangibles in firms' capital stocks, as predicted by theory.[9]

Conducting a natural field experiment is challenging in this situation because US regulators have been heretofore unwilling to randomize accounting policies to different firms. As we discuss later, regulators are softening their stance toward experiments, but there is still work to be done before accounting researchers can work with the Financial Accounting Standards Board (FASB), Public Company Accounting Oversight Board (PCAOB), and Securities and Exchange Commission (SEC) to randomize disclosure standards. One potential avenue for future

[9] Subgroup analysis can be particularly useful for providing evidence on the mechanism driving the treatment effect.

implementation, however, is to utilize a “phase-in” period in which a random set of firms adheres to the new rule before it is implemented more broadly. To conduct appropriate analyses, the FASB and SEC would need to require firms to submit intangible expenses to the research group, regardless of whether the firm was required to publicly disclose intangible expenses. Alternatively, researchers may be able to address data availability problems by collecting data directly from firms.

An additional way to implement the experiment is to conduct a natural field experiment in countries outside of the US (or conduct framed and artefactual field experiments).[10] Developing economies can be appealing because they often consist of firms that are substantially smaller in scale, and thus concerns about negative effects are more muted. For example, a growing literature in economics has utilized India as an environment to conduct field experiments. Of course, the limitation of conducting field experiments in developing economies mirrors the concerns regarding analyzing real effects in restaurant and MBA labor markets—generalizability. This is a limitation of all studies, including studies within the US that implicitly generalize outside of the sample period being studied. We discuss this further in the remaining sections of the paper.

Despite its conceptual appeal, the skeptical accounting researcher may remain unconvinced that such a field experiment can be successfully implemented. However, economists faced a similar issue two decades ago when many thought that the process of experimentation was best left to work in the lab. Since then, economists have developed myriad creative approaches that have allowed them to draw strong inferences using the field

[10] Later, we discuss how artefactual field experiments may be important first steps for conducting field experiments with regulators and firms.

experimental approach. The goal of our discussion is to encourage accounting researchers to do the same.

The remainder of this section utilizes fields in economics that are conceptually similar to accounting to detail how economists have conducted field experiments and how they might provide useful insights and ideas for future accounting research.

Auditing

The first setting we discuss is auditing and third-party verification markets. Inference in the auditing literature is challenging because lack of exogenous variation limits researchers' ability to answer key theoretical questions (Minnis [2011]). Auditors are not randomly assigned to firms as in experiments, thereby making it difficult to draw causal conclusions for various research questions (e.g., audit quality). Furthermore, the empirical constructs of interest are often difficult to measure. DeFond and Zhang [2014] provide an extensive discussion of the myriad issues in measuring audit quality. We argue, at least to some extent, that this problem can be addressed by field experiments, which, as data-generating processes, give the researcher control over how economic concepts are measured.

We discuss a study (Duflo et al. [2013; DGPR hereafter]) that conducts a field experiment in Gujarat, India, to address third-party auditing issues involving pollution. Pollution auditing is not a traditional accounting environment, but the parallels between the two settings afford us the opportunity to discuss several important issues. We return to the issue of generalizability to other accounting settings at the conclusion of this section.

DGPR conduct a randomized auditing treatment for polluting firms in India. The status quo in India prior to the study is characterized by both significant pollution and corruption in reporting of pollution levels to regulators. The DGPR setting shares features with financial accounting markets in the United States because firms are required to select auditors and pay them with their own funds. The experimental treatment in DPGR consists of four components that impact the auditing process: random auditor selection, payment from a central pool of funds, random back-checking, and a payment for performance provision. The goal of the treatment is to increase auditor independence and therefore the incentives to report truthfully. The increased transparency in the audit report could also have the real effect of reducing actual pollution levels. Importantly, these four treatments are implemented simultaneously, so the authors cannot draw strong conclusions concerning which channel (or subsets of channels) the treatment operates through.

The authors find two empirical facts that are of interest to accounting researchers, using two primary outcome variables: reported pollution levels and back-checked pollution readings collected by the authors. First, firms not subject to the treatment report pollution levels that are directly below the regulatory pollution threshold at abnormally high frequencies. This is reminiscent of accounting earnings that are seen with abnormal density immediately above analyst forecast targets and other profitability thresholds (Degeorge, Patel, and Zeckhauser [1999]). Second, treatment has a statistical and economic impact on the reporting decisions of firms. Firms in the treatment group do not exhibit the same tendency to report pollution outcomes directly below the regulatory threshold, which provides evidence that auditor independence leads to more truthful reporting. Furthermore, the paper also provides evidence of real effects that result from higher-quality reporting by illustrating that firms reduce their pollution levels in response to the treatment.

DGPR alleviate identification concerns by randomly assigning treatment to firms. Randomization allows them to demonstrate the attainment of covariate balance between control and treatment prior to experimental intervention. This, in turn, allows them to draw causal interpretations concerning auditor independence and incentive designs that are difficult to make in the existing archival literature.

The design DGPR use avoids measurement problems by directly measuring audit quality. Their ability to provide a convincing measure of audit quality is a direct benefit of conducting a field experiment, since the researchers control the data-generating process. Consequently, they design the experiment in order to collect a precise measure of the economic forces they are interested in. This often contrasts with archival research in which the variables used for analyses are exogenous to the researcher, and may not directly serve the needs of the research. DGPR back-check pollution levels with independently hired industry experts. These back-checks provide a convincing measure of true pollution levels that can be compared to reported pollution levels, thereby accounting for the underlying quality of the firm. Although this measure is subject to its own criticisms, it provides a useful measure of the degree to which auditors deviate from truthful values—in other words, audit quality that is directly related to the incentives of auditors to alter reported outcomes.

An additional advantage of the setting in DGPR is that the experiment is conducted in a real-world auditing environment. It would be difficult, if not impossible, to recreate many of the conditions of a real-world auditing environment within a lab. A limitation of the pollution auditing setting, however, is that it may not generalize to other accounting settings, such as financial accounting for publicly listed US firms. Although field experiments typically improve on generalizability relative to lab experiments, they are not a panacea. Economic theory helps to determine the extent to which the results of studies, such as DGPR's, can be generalized to other markets. For example, the US has more complex regulatory institutions and different market structures for auditing products, both of which imply that the results of a pollution auditing study in India may not be perfectly informative to US financial audits. Conducting a field experiment in the US would alleviate this problem. However, until such a study is conducted, theory can help us understand how we expect the treatment effect to differ if such a policy were to be implemented in the US.

Field experiments conducted in settings that are not directly applicable to the primary market of interest moderate the degree to which researchers might update their priors. This framework allows for a more flexible approach for interpreting the results of DGPR in contrast to simply classifying the study as "relevant" or "not relevant" to accounting researchers. Imagine a policymaker with the explicit goal of increasing auditor independence and truthful reporting in the US who is considering the treatment addressed in DGPR. Although the policymaker would prefer a more direct study with respect to her setting, DGPR still allows her to adjust her prior in such a way that she makes more informed policy decisions.

Finally, we note that opportunities for field experiments in auditing within the US are beginning to emerge. The PCAOB, Center for Economic Analysis has an explicit goal of using economic theory and analysis to inform PCAOB activities (PCAOB [2015]). As exemplified by the joint PCAOB/Journal of Accounting Research conferences, the Center for Economic Analysis has demonstrated interest in working with and exchanging dialog with accounting and finance academics. As this relationship grows, we envision an opportunity for the PCAOB to work with academics in order to run policy-relevant randomized field studies.

Financial Reporting and Measurement

A substantial subset of accounting research is focused on the need to reliably measure firm performance and subsequently report performance to external stakeholders. Therefore, insights into the tradeoffs of different measurement systems are imperative to understanding the implications of accounting choice. Field experiments are a potential methodology to expand knowledge in this area.

This line of research has historically conducted studies that fall in the artefactual field experiment category. These studies include early work (e.g., Barrett [1971]) that explores the valuation consequences of different financial statement presentations. To do so, existing research has utilized subject pools of financial analysts.

Working with the SEC and the FASB to implement framed field experiments, however, is an appealing alternative. This is the most direct avenue in which researchers can test the capital market effects of different measurement systems, including different accounting standards. However, convincing regulatory bodies to implement an experiment that requires substantially different accounting policies to be randomized across firms is challenging. Regulators face real-world constraints, such as the presence of political costs, which limit their ability to engage in large-scale experimentation.

Despite this concern, recent developments give reason for optimism. Experiments such as the Regulation SHO Pilot Program have been implemented in financial settings. The SEC instituted the pilot program to better understand short selling behavior in 2005. More recently, the SEC approved a pilot program to understand the impact of wider tick sizes on the market quality of stocks for smaller companies (SEC [2015]). These programs suggest the circumstances that researchers should focus on when anticipating whether regulatory bodies are willing to conduct experimentation. On one hand, conducting a large-scale field experiment needs to be justified from the perspective of generating substantial economic insight. On the other hand, the experiment cannot produce financial or political costs that outweigh the benefits from developing these insights.

Although convincing regulators to directly intervene in financial markets is the most conceptually appealing strategy for many research questions, there are other avenues in which field experiments can contribute to financial reporting research. Determining market values for goods that are not traded on active markets is a crucial issue for policymaking. Landry and List [2007] is one example of a paper that investigates this issue. The authors explore the degree to which hypothetical willingness to pay differs from actual willingness to pay (i.e., payments) in the sports card market. The implications of this are important for determining whether contingent valuation is an accurate mechanism for determining market prices. Though the results of the study are interesting in their own right (they find a significant difference between hypothetical and real values), of interest to accounting researchers is the degree to which contingent valuation faces similar challenges as fair value accounting: valuing non-market assets. Once again, the insights for accounting rely on how dependent the findings in Landry and List [2007] are on the particular setting in which it was conducted (in this instance, they may be highly dependent). Though, our point is to suggest that formulating theory and testing these theories in comparable settings could contribute substantial economic insight to financial reporting research.[11]

Firm Disclosure

[11] For example, List [2006] uses the sports card market to address how social preferences and third-party verification influence market outcomes. Ultimately, List [2006] is interested in understanding market behavior that extends beyond the sports card market. List [2006] shows that researchers can use real-world market settings that are not explicitly the setting of interest to develop important insights.

Researchers can also take a broader perspective on what constitutes accounting information to include a diverse set of firm disclosure channels that extend beyond financial statements and SEC filings. Examples in this area include press releases, conference calls, and analyst reports.

Much of the discussion in this area of research mirrors previous sections. However, the questions asked in the broader disclosure literature allow the researcher to utilize even more diverse settings. Bursztyn et al. [2014] capitalize on a partnership with a financial brokerage firm to conduct a field experiment relating to the demand for investment products. The paper is framed as a study in the peer effects literature, but it can also interpreted as a field experiment in disclosure. In their experiment, the researchers examine the effects of presenting information to consumers about the investment behavior of their peers on consumer demand for a particular investment option. Their primary result is that peers influence asset purchasing decisions through both learning and social utility mechanisms (peers internalize value for owning similar assets).

There is a wealth of information channels that emanate from the firm that are in need of more rigorous understanding. Heinrichs [2015] uses a field experiment to better understand private information channels between management and investors. She finds evidence that this channel is economically significant (58% of firms respond to emails seeking access to corporate management). Furthermore, she shows that firms responded differently to emails depending on the party requesting management access. She finds an increased likelihood of management to respond to emails from domain names that reference investing relative to consulting.

When considering alternative information channels, one can also utilize alternative treatment mechanisms. For example, consider Banerjee et al. [2015], who conduct a field experiment in which they investigate whether disseminating information regarding social programs increases citizen participation. The researchers implement their experiment by mailing information about the program, designed by the experimenters, to citizens, who otherwise would need to find information about the program by accessing government sources. Researchers in accounting might consider similar mechanisms for disseminating financial information that would allow for the exploration of the consequences of conveying accounting information or other firm disclosures to investors.

Managerial Accounting

The accounting literature has also researched the internal use of accounting information for firm decision making (e.g., Ittner and Larker [2001]). For ease of discussion, we partition this area into two (related) subsections: (1) the evaluation, compensation, and incentives of employees (entry-level employees to top executives) and (2) the implementation of control systems used for internal decision making.

Incentive Structures

A perusal of the incentives literature reveals that prior research has paid considerable attention to the compensation and incentives of top executives. This contrasts with studies that look at incentives across a broader distribution of employees within the firm (e.g. Hochberg and Lindsey [2010]). Yet the compensation and incentives of both groups are important considerations within organizational design and accounting research.

From a researcher's perspective, natural field experiments with the explicit goal of understanding executive compensation are the most difficult to execute. Framed and artefactual field experiments, on the other hand, are a more realistic opportunity for most researchers conducting experimental-based inference in this area. For example, similar to Graham, Harvey, and Rajgopal [2005] and Fehr and List [2004], there are opportunities to partner with organizers of financial executive conferences to include high-level managers in the subject pools of studies that are interested in understanding employee incentives. Importantly, even if a researcher successfully conducts such an experiment, researchers must be aware that because these are often not natural experiments, as discussed above, there is potential for selection effects through the differential inclusion of managers who are willing to participate, thereby limiting the generalizability of the experiment.

One example of an artefactual field experiment that utilizes subjects in leadership positions is Kosfeld and Rustagi [2015]. The study conducts an experiment using leaders of forest commons in Ethiopia. The authors intend to address how different leadership styles affect success within group behavior. They find that leaders who promote equality and efficiency see better outcomes than leaders who punish without careful consideration.[12]

There is considerable opportunity for field experiments that utilize subject pools other than leaders and executives. In this area, personnel economics has utilized field experiments to better understand employee incentive structures. Bandiera, Barankay, and Rasul [2011] review some of the important contributions in this area. Notably, the authors discuss the work of Shearer [2004], which examines the productivity effects of changing the structure of workers' monetary incentives. Shearer [2004] conducts a field experiment through a partnership with a tree-planting firm in British Columbia. He finds that moving from a fixed-wage to a piece-rate compensation scheme increases worker productivity by approximately 20%.

[12] In contrast to studies that focus on the incentives of CEO behavior, the paper demonstrates the consequences of leaders' characteristics. This may be relevant to the growing literature interested in the relationship between CEO personality and decision making (e.g., Gow et al. [2015]).

The lessons from Shearer [2004] reinforce a theme that is especially relevant to researchers in accounting: partnering directly with organizations can be an effective strategy. Partnering with organizations that have multiple business units is appealing for many reasons, such as statistical power.[13] Smaller-scale partnerships can still be promising, as well. Kube, Maréchal, and Puppe [2013] partner with a university library in order to show that wage cuts substantially reduce productivity (wage increases, however, do not result in productivity gains).

Employee theft is of potential interest to accounting researchers trying to understand the incentives and behaviors of employees within the firm that extend beyond productivity (e.g., Chen and Sandino [2012] in the accounting literature). Gill, Prowse, and Vlassopoulos [2013] conduct an online experiment that examines the effects of bonuses on productivity. They find that compensation based on random bonuses encourages cheating within the workplace. Nagin et al. [2002] design a field experiment in which they test the predictions of the rational cheater model. They find that experimentally lowering the perceived cost of cheating for telephone solicitation employees increases the rate at which they cheat. However, this primarily occurs for employees who have negative perceptions of their employers; the remaining employees are insensitive to changes in monitoring. Field experiments are an important possibility for developing research that extends beyond productivity, especially given the potential for experimenter demand effects to suppress morally sensitive behaviors in the lab (Levitt and List [2007]).

Internal Controls and Management Best Practices

[13] If the researcher partners with organizations such as Starbucks, for example, then this affords a potentially large number of observations to detect a treatment effect (if randomization occurred at the coffee-shop level).

Understanding employee incentives in accounting settings is part of the broader theme of understanding and designing proper control systems within an organization. Control systems can be thought of as establishing the proper targets and goals of the firm while concurrently implementing best practices in how to measure and achieve them. The elements of implementing a proper control system can range from the proper measurement of internal performance to the strategies and processes that utilize the information. To a large extent, existing research utilizes case study methods (Scapens [1990]). Although internally valid, this approach significantly limits the generalizability of results. Alleviating this concern would likely require partnering with multiple firms or utilizing firms with multiple operating segments, as discussed previously. Given the challenges that this often presents, we propose another possibility. Researchers have increasingly utilized partnerships with third-party organizations in order to apply treatment to multiple firms. For example, Bloom et al. [2013] find that assigning management consultants to different firms can have economically significant impacts on firm productivity. We envision opportunities for relationships with third parties in order to help determine best practices in accounting within organizations.

As a final example, Bloom et al. [2015] conduct a field experiment in China to determine the effects of organizational structure on performance. The authors address the extent to which working from home either increases or decreases performance in a Chinese travel agency. Interestingly, they find that working from home leads to a 13% improvement in performance relative to a randomly assigned control group. They also document an increase in job satisfaction. The study highlights two important features. They are able to investigate a question that is almost inaccessible without utilizing field experiments. Second, the paper reinforces that field experiments can provide important evidence on determining management best practices, a critical topic in managerial accounting research.

International Accounting

A large body of accounting research has examined the use of accounting information in developed economies. As an example, within these economies, debt markets have been of considerable interest to both accounting and finance researchers. Constructing a field experiment that investigates debt markets within developed financial markets is a particularly potent example of an issue raised throughout this discussion. Because of regulatory scrutiny, it is difficult (although not impossible) to form a partnership with a firm or third-party organization that would allow the random manipulation of important elements of debt contracts within US financial markets.

A growing literature in economics looks at debt markets in a diverse set of financial markets (including developing economies). Several papers, including Field and Pande [2008] and Feigenberg, Field, and Pande [2013], use microfinance as a setting to examine various aspects of debt financing. Feigenberg, Field, and Pande [2013] look at the effects of social interaction on loan default and participants' willingness to share risk in Kolkata. Similarly, using microfinance firms in developing countries may be a useful setting for accounting researchers to understand the effects of changes in accounting constructs (e.g.. accounting quality). Pérez Cavazos [2015] is a fitting example of an archival study in accounting that uses a microfinance setting to investigate how borrower communication impacts strategic default.

We additionally note that critiques similar to those discussed throughout this paper apply to the microfinance setting. One large concern is that the features of microfinance do not generalize to developed markets, such as the US. Researchers should be aware of the differences between settings such as microfinance in developing markets and the US market so that they can carefully characterize how relevant their findings are to traditional accounting settings. It is worth mentioning, however, that understanding accounting in microfinance (and other developing economy settings) is potentially interesting in its own right.

Despite this limitation, the use of developing economies is appealing for several reasons. First, as noted above, the potential for treatment randomization is potentially higher due to the political climate and smaller scale in which these economies operate. Second, an understanding of developing economies can help researchers better understand developed economies by investigating the processes and forces in which they evolve. When examining developed financial markets, it is often difficult to disentangle preferences that guide individual decision making from the complexities of the setting in which agents operate. By looking at a setting with limited complexities, the researcher has greater potential to examine the underlying decision-making processes that impact financial markets.

Tax

The study of tax in accounting research often focuses on the institutional effects of tax law (see Hanlon and Heitzman [2010] for a review of research in this area). Within this field, tax avoidance is an area that is particularly well suited for field experiments.

Choo, Fonseca, and Myles [2014] conduct an experiment to determine the incentives for tax compliance. They find that compliance is dependent on the severity of the financial penalty. However, they find no evidence that an increase in audit probability has any substantial effect on compliance. This evidence provides some guidance for researchers who are interested in understanding IRS enforcement.

The study includes subject pools that are representative of real-world populations. Interestingly, the paper demonstrates that this sample selection matters. Students in the experiment behave differently from company employees and self-employed taxpayers. The authors attribute this difference to norms that exist within the field that are not present for student subjects. Consequently, the paper demonstrates the importance of looking at institutional issues, such as tax compliance, within the field. Ignoring these particular institutional norms could lead the researcher to draw substantially different conclusions concerning the effects of tax policy.

9. What Makes Field Experiments in Accounting and Finance Challenging?

Despite our belief that field experiments have not been utilized as effectively as possible in accounting and finance, we recognize that there are challenges that exist in terms of implementation. These barriers to entry are important to discuss and outline so future scholars understand the nature of the field they are entering. We summarize a few of the major concerns in this section.

Accounting research is often focused on settings where randomization may be difficult to achieve: Our discussion above demonstrates that field experiments are already being conducted in fields such as auditing and management accounting research, which both exist under the umbrella of research interesting to accounting professors. However, a large proportion of researchers are interested in understanding how accounting disclosure affects participants in capital markets. Randomization, in this setting, may be more difficult to achieve at the firm level.

However, advancements in technologies and a deeper understanding of situations conducive to experimental explorations will surely open doors in this area.

Field experiments in accounting will often require participation from firms, which may be unwilling to participate in acts of randomization: Accounting is institutional in nature; thus, partnering with firms is essential when conducting some types of field experiments. In our experience, many firms may be hesitant to implement randomization because of concerns regarding the consequences of being perceived as unfair. In this way, one useful first step is to use managers and decision makers in artefactual and framed field experiments that permit a first glimpse at how variables are related. This allows the decision makers and firms to realize what is possible, as well as the potential gains to experimentation. After this, natural field experiments will naturally arise.

Field experiments are costly to the researcher to execute and implement: The conventional wisdom is that field experiments are costly for (at least) two reasons. First, field experiments can be monetarily costly. Some field experimental costs will greatly exceed researchers' existing budgets, requiring researchers to apply for funding. Second, field experiments often require many years to prepare and implement, and they may also require longitudinal data to explore the underlying questions of interest. In this way, the long window from design to publication may be inconsistent with researchers' incentives with respect to tenure clocks.

In terms of the first consideration—cost—in our experience, we have learned that many field experiments can be naturally conducted in the everyday course of business. In this way, with a little thought and creativity, field experiments can be free, in that there are no new outlays. This line of thinking reverses the cost argument to the following: it is costly for the firm *not* to experiment because they are foregoing learning opportunities daily by not knowing what works in their business and why. Opportunities naturally arise if you look hard enough.

The second consideration—timely investment—is that field experiments with firms or regulators often require long-term investment. Engaging a firm with the idea of randomization—without a preexisting relationship—is ambitious. We stress that the most effective process requires a researcher to build a foundation of research and relationships from which a field experiment can emerge. For example, a researcher might consider partnering with firms to solve problems that the firm is currently facing and in the process conduct field study research with the aim of potentially conducting a field experiment when the opportunity arises and trust has been established between the researcher and organization. Additionally, as discussed above, researchers may begin their research portfolios by conducting artefactual field experiments, which deviate from lab experiments by utilizing non-standard subject pools and are therefore more practical to implement in the short run. If researchers can demonstrate the value of research and randomization using artefactual field experiments and field studies, firms and regulators will increasingly become interested in implementing randomization to their own benefit.

Build It and They will Come

As the astute reader likely realizes, many of the concerns outlined above are met with a response from us that is akin to "build it and they will come," meaning that once the operation of field experiments starts, opportunities that not even the imagination can envision will indeed arise. More than 20 years ago, people were skeptical of such a rise in economics, but it has happened.

To more formally address the process we have in mind, we draw on an example from our experience in conducting field experiments over the past 20+ years. For instance, consider the line of research in finance called myopic loss aversion (MLA). MLA is a theory that Benartzi and Thaler [1995] outline that combines two behavioral concepts—loss aversion (Kahneman and Tversky [1979]) and mental accounting (Thaler [1985])—to provide a theoretical foundation for the observed equity premium puzzle in finance. Early on, lab experiments with students were consistent with individual behavior following the MLA model (see, e.g., Gneezy and Potters [1997] and Gneezy, Kapteyn, and Potters [2003]). Haigh and List [2005] extend this research to market professionals by using a framed field experiment to test whether professional futures and options pit traders from the Chicago Board of Trade behave similarly. They report evidence in the affirmative; indeed, professional traders exhibit behavior consistent with MLA to a *greater* extent than undergraduate students.

In this way, the “it” had been built through lab and a framed field experiment. Having read this line of research, Francis Larson, founder of Normann, reached out to List and his colleagues to ask if they were interested in experimenting with his firm. Normann could provide a trading platform for retail foreign exchange (FX) traders that would allow a test of MLA using a natural field experiment. Never imagining this was possible, List took the opportunity, and MLA theory is now being tested in a natural field experiment, with Larson, List, and Robert Metcalfe as coauthors.

We would not find it difficult to discuss several other lines of research of which we have engaged in economics, accounting, and finance that have followed similar paths. For instance, our work exploring optimal approaches to induce tax non-compliers to pay their tax bills has now become a series of natural field experiments with the UK behavioral insights team. Our work on the economics of charity began with a set of artefactual field experiments and has now become a set of natural field experiments with dozens of charities. Likewise, energy conservation and work more broadly in environmental and resource economics have become a series of natural field experiments with OPower and Virgin Atlantic Airlines. Our line of research in labor economics has expanded to include participation from myriad sources, from Chrysler to many school districts around the US to factory workers in Chinese manufacturing plants. All of these, as well as many more examples that we have left on the sidelines, showcase that if you build it, they will come. We very much look forward to watching this transformation in accounting and finance.

10. Concluding Thoughts

An empirical revolution has arrived in the social sciences. This revolution has been important in that the credibility of received empirical estimates has increased considerably. The revolution began more than 30 years ago in economics, as scholars began to take the modeling of causal relationships in naturally occurring data to new levels. Whether using a difference-in-differences approach or instrumental variables, working diligently to identify clean treatment effects has become more the norm than the exception.

In the last two decades, a new type of empirical approach has taken hold in economics: field experiments. Field experiments use randomization to identify treatment effects and in many cases do so in a realistic environment. In this way, field experiments have the best of both worlds: an approach to identify causal relationships that relies on minimal assumptions (randomization) performed in the markets of ultimate interest.

In this study, we summarize the recent literature on field experiments in economics in a manner that lends itself well to future applications in accounting and finance. Beyond outlining the "hows" and "whys" of experimentation—from partnering with firms to the nuts and bolts of design to data analysis to building scientific knowledge—we summarize areas in accounting and finance that are ripe for field experiments. We view current practices within firms as largely a black box—an important box but black nonetheless. Field experimentation represents a powerful tool to open up secrets within that box, and when combined with applicable theory, it can provide a deeper understanding heretofore unachieved. That deeper understanding represents the promise of field experimentation. We hope that, in some small way, our study can help field experiments reach their potential in accounting and finance.

References

ALEVY, J. E., M. S. HAIGH, and J. A. List. ‘Information Cascades: Evidence from a Field Experiment with Financial Market Professionals.’ The Journal of Finance 62 (2007): 151-180.

AL-UBAYDLI, O., and J. A. List. On the Generalizability of Experimental Results in Economics. Methods of Modern Experimental Economics, Oxford University Press, 2013.

AL-UBAYDLI, O., and J. A. LIST. ‘Do Natural Field Experiments Afford Researchers More or Less Control than Laboratory Experiments?’ American Economic Review 105 (2015): 462-66.

ANDERSON, M. ‘Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.’ Journal of the American Statistical Association 103 (2008): 1481-95.

BANDIERA, O., I. BARANKAY, and I. RASUL. ‘Field Experiments with Firms.’ The Journal of Economic Perspectives 25 (2011): 63-82.

BANERJEE, A., R. HANNA, J. C. KYLE, B. A. OLKEN, and S. SUMARTO. ‘The Power of Transparency: Information, Identification Cards and Food Subsidy Programs in Indonesia.’ Working paper, 2015.

BARRETT, M. E. ‘Accounting for Intercorporate Investments: A Behavior Field Experiment.’ Journal of Accounting Research 9 (1971): 50-65.

BENARTZI, S., and R. H. THALER. ‘Myopic Loss Aversion and the Equity Premium Puzzle.’ The Quarterly Journal of Economics 110 (1995): 73-92.

BLOOM, N., B. EIFERT, A. MAHAJAN, D. MCKENZIE, and J. ROBERTS. ‘Does Management Matter? Evidence from India.’ The Quarterly Journal of Economics 128 (2013): 1-51.

BLOOM, N., J. LIANG, J. ROBERTS, and Z. J. YING. ‘Does Working from Home Work? Evidence from a Chinese Experiment.’ The Quarterly Journal of Economics 130 (2015): 165-218.

BLOOMFIELD, R., M. W. NELSON, and E. SOLTES. ‘Gathering Data for Financial Reporting Research.’ Working paper, 2015.

BLUNDELL, R., and M. COSTA DIAS. ‘Alternative Approaches to Evaluation in Empirical Microeconomics.’ Portuguese Economic Journal 1 (2002): 91-115.

BONFERRONI, C. E. ‘Il Calcolo delle Assicurazioni su Gruppi di Teste.’ Studi in Onore del Professore Salvatore Ortu Carboni (1935): 13-60.

BURSZTYN, L., F. EDERER, B. FERMAN, and N. YUCHTMAN. ‘ Understanding Mechanisms Underlying Peer Effects: Evidence from a Field Experiment on Financial Decisions.’ Econometrica 82 (2014): 1273-1301.

CHEN, C. X., and T. SANDINO. ‘Can Wages Buy Honesty? The Relationship between Relative Wages and Employee Theft.’ Journal of Accounting Research 50 (2012): 967-1000.

CHOO, L., M. A. FONSECA, and G. D. MYLES. ‘Do Students Behave Like Real Taxpayers? Experimental Evidence on Taxpayer Compliance from the Lab and from the Field.’ Tax Administration Research Center, 2014.

CHRISTENSEN, H., E. FLOYD, L. YAO LIU, and M. MAFFETT. ‘The Real Effects of Mandatory Non-Financial Disclosures in Financial Statements.’ Working paper, 2015.

CHRISTENSEN, H., L. HAIL, and C. LEUZ. ‘Capital-Market Effects of Securities Regulation: Prior Conditions, Implementation, and Enforcement.’ Working paper, 2015.

DEFOND, M., and J. ZHANG. ‘A Review of Archival Auditing Research.’ Journal of Accounting and Economics 58 (2014): 275-326.

DEGEORGE, F., J. PATEL, and R. ZECKHAUSER. ‘Earnings Management to Exceed Thresholds.’ Journal of Business 72 (1999): 1-33.

DELLAVIGNA, S., J. A. LIST, and U. MALMENDIER. ‘Testing for Altruism and Social Pressure in Charitable Giving.’ The Quarterly Journal of Economics 127 (2012): 1-56.

DELLAVIGNA, S., J. A. LIST, and U. MALMENDIER. ‘Voting to Tell Others.’ Working paper, 2015.

DUFLO, E., M. GREENSTONE, R. PANDE, and N. RYAN. ‘Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India.’ The Quarterly Journal of Economics 128 (2013): 1499-1545.

FEHR, E., and J. A. LIST. ‘The Hidden Costs and Returns of Incentives – Trust and Trustworthiness among CEOs.’ Journal of the European Economic Association 2 (2004): 743-771.

FEIGENBERG, B., E. FIELD, and R. PANDE. ‘The Economic Returns to Social Interaction: Experimental Evidence from Microfinance.’ The Review of Economic Studies 80 (2013): 1459-1483.

FIELD, E., and R. PANDE. ‘Repayment Frequency and Default in Microfinance: Evidence from India.’ Journal of the European Economic Association 6 (2008): 501-509.

FINK, G., M. MCCONNELL, and S. VOLLMER. ‘Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures.’ Journal of Development Effectiveness 6 (2014): 44-57.

FLOYD, E., and S. TOMAR. ‘The Effects of Disclosure Policy on Human Capital Investment Signaling: The Case of MBA Grade Non-Disclosure.’ Working paper, 2015.

GILL, D., V. PROWSE, and M. VLASSOPOULOS. ‘Cheating in the Workplace: An Experimental Study of the Impact of Bonuses and Productivity.’ Journal of Economic Behavior & Organization 96 (2013): 120-134.

GNEEZY, U., and J. POTTERS. ‘An Experiment on Risk Taking and Evaluation Periods.’ The Quarterly Journal of Economics 112 (1997): 631-645.

GNEEZY, U., A. KAPTEYN, and J. POTTERS. ‘Evaluation Periods and Asset Prices in a Market Experiment.’ Journal Title 58 (2003): 821-838.

GOW, I. D., S. N. KAPLAN, D. F. LARCKER, and A. ZAKOLYUKINA. ‘Executive Personality.’ Working paper, 2015.

GOW, I. D., D. F. LARCKER, and P. C. REISS. ‘Causal Inference in Accounting Research.’ Working paper, 2015.

GRAHAM, J. R., C. R. HARVEY, and S. RAJGOPAL. ‘The Economic Implications of Corporate Financial Reporting.’ Journal of Accounting and Economics 40 (2005): 3-73.

HAIGH, M. S., and J. A. LIST. ‘Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis.’ Journal of Finance 60 (2005): 523-534.

HALLSWORTH, M., J. A. LIST, and R. METCALFE. ‘The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance.’ Working paper, 2014.

HANLON, M., and S. HEITZMAN. ‘A Review of Tax Research.’ Journal of Accounting and Economics 50 (2010): 127-178.

HARRISON, G. W., and J. A. LIST. ‘Field Experiments.’ Journal of Economic Literature 42 (2004): 1009-1055.

HECKMAN, J. J. ‘The Scientific Model of Causality.’ Sociological Methodology 35 (2005): 1-97.

HEINRICHS, A. ‘Investors’ Access to Corporate Management: A Field Experiment about 1-on-1 Calls.’ Working paper, 2015.

HOCHBERG, Y., and L. LINDSEY. ‘Incentives, Targeting, and Firm Performance: An Analysis of Non-executive Stock Options.’ Review of Financial Studies 23 (2010): 4148-4186.

HOLM, S. ‘A Simple Sequentially Rejective Multiple Test Procedure.’ Scandinavian Journal of Statistics 6 (1979): 65-70.

HOSSAIN, T., and J. MORGAN. ‘Plus Shipping and Handling: Revenue (Non)-Equivalence in Field Experiments on eBay.’ Advances in Economic Analysis & Policy 6 (2006): 3.

ITTNER, C. D., and D. F. LARCKER. ‘Assessing Empirical Research in Managerial Accounting: A Value-based Management Perspective.’ Journal of Accounting and Economics 32 (2001): 349-410.

JIN, G. Z., and P. LESLIE. ‘The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards.’ The Quarterly Journal of Economics 118 (2003): 409-451.

KAHNEMAN, D., and A. TVERSKY. ‘Prospect Theory: An Analysis of Decision under Risk.’ Econometrica 47 (1979): 263-292.

KANODIA, C., and H. SAPRA. ‘A Real Effects Perspective to Accounting Measurement and Disclosure: Implications and Insights for Future Research.’ Working paper, 2015.

KANODIA, C., H. SAPRA, R. VENUGOPALAN. ‘Should Intangibles Be Measured: What Are the Economic Trade-Offs?’ Journal of Accounting Research 42 (2004): 89-120.

KOSFELD, M., and D. RUSTAGI. ‘Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia.’ The American Economic Review 105 (2015): 747-783.

KUBE, S., M. A. MARÉCHAL, and C. PUPPE. ‘Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment.’ Journal of the European Economic Association 11 (2013): 853-870.

LANDRY, C. E., and J. A. LIST. ‘Using Ex Ante Approaches to Obtain Credible Signals for Value in Contingent Markets: Evidence from the Field.’ American Journal of Agricultural Economics 89 (2007): 420-429.

LEUZ, C., and P. WYOSOCKI. ‘The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research.’ Working paper, 2015.

LEVITT, S. D., and J. A. LIST. ‘What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?’ The Journal of Economic Perspectives 21 (2007): 153-174.

LEVITT, S. D., and J. A. LIST. ‘Field Experiments in Economics: The Past, the Present, and the Future.’ European Economic Review 53 (2009): 1-18.

LIBBY, R., R. BLOOMFIELD, and M. W. NELSON. ‘Experimental Research in Financial Accounting.’ Accounting, Organizations, and Society 27 (2002): 775-810.

LIBBY, R., and T. BROWN. ‘Financial Statement Disaggregation Decisions and Auditors’ Tolerance for Misstatement.’ The Accounting Review 88 (2013): 641-665.

LIST, J. A. ‘Does Market Experience Eliminate Market Anomalies?’ The Quarterly Journal of Economics 118 (2003): 41-71.

LIST, J. A. Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace.’ Econometrics 72 (2004): 615-625.

LIST, J. A. ‘Testing Neoclassical Competitive Theory in Multilateral Decentralized Markets.’ Journal of Political Economy 112 (2004): 1131-1156.

LIST, J. A. ‘Field Experiments: A Bridge between Lab and Natural Occurring Data.’ The B.E. Journal of Economic Analysis and Policy 5 (2006): 8.

LIST, J. A. ‘The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions.’ Journal of Political Economy 114 (2006): 1-37.

LIST, J.A. ‘Informed Consent in Social Science.’ Science 322 (2008): 672.

LIST, J. A. ‘Does Market Experience Eliminate Market Anomalies? The Case of Exogenous Market Experience.’ American Economic Review 101 (2011): 313-317.

LIST, J. A. ‘Why Economists should Conduct Field Experiments and 14 Tips for Pulling One Off.’ Journal of Economic Perspectives 25 (2011): 3-15.

LIST, J. A., and RASUL, I. ‘Field Experiments in Labor Economics.’ Handbook of Labor Economics 4 (2011): 103-228.

LIST, J. A., S. SADOFF, and M. WAGNER. ‘So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design.’ Experimental Economics 14 (2011): 439-457.

LIST, J. A., A. M. SHAIKH, and Y. XU. ‘Multiple Hypothesis Testing in Experimental Economics.’ Working paper, 2015.

MANIADIS, Z., F. TUFANO, and J. A. LIST. ‘One Swallow Doesn’t Make a Summer: New Evidence on Anchoring Effects.’ The American Economic Review 104 (2014): 277-290.

MINNIS, M. ‘The Value of Financial Statement Verification in Debt Financing: Evidence from Private U.S. Firms.’ Journal of Accounting Research 49 (2011): 457-506.

NAGIN, D., J. REBITZER, S. SANDERS, and L. TAYLOR. ‘Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment.’ American Economic Review 92 (2002): 850-73

PUBLIC COMPANY ACCOUNTING OVERSIGHT BOARD (PCAOB). Center for Economic Analysis. Retrieved from: http://pcaobus.org/About/CenterforEconomicAnalysis/Pages/default.aspx, 2015.

PÉREZ CAVAZOS, G. ‘Sharing Private Information with Customers: Strategic Default and Lender Learning.’ Working paper, 2015.

ROSENBAUM, P. R., and D. B. RUBIN. ‘The Central Role of the Propensity Score in Observational Studies for Causal Effects.’ Biometrika 70 (1983): 41-55.

ROSENZWEIG, M. R., and K. I. WOLPIN. ‘Natural “Natural Experiments” in Economics.’ Journal of Economic Literature 38 (2000): 827-874.

SCAPENS, R. W. ‘Researching Management Accounting Practice: The Role of Case Study Methods.’ The British Accounting Review 22 (1990): 259-281.

SECURITIES AND EXCHANGE COMMISSION (SEC). SEC Approves Pilot to Assess Tick Size Impact for Smaller Companies. Retrieved from: http://www.sec.gov/news/pressrelease/2015-82.html, 2015.

SHEARER, B. ‘Piece Rates, Fixed Wages, and Incentives: Evidence from a Field Experiment.’ Review of Economic Studies 71 (2004): 513-534.

SMITH, V. L. ‘Relevance of Laboratory Experiments to Testing Resource Allocation Theory.’ Evaluation of Econometric Models, 1980.

SMITH, V. L. ‘Microeconomic Systems as an Experimental Science.’ American Economic Review 72 (1982): 923-55.

THALER, R. H. ‘Mental Accounting and Consumer Choice.’ Marketing Science 4 (1985): 199-214.

WACHOLDER, S., S. CHANOCK, M. GARCIA-CLOSAS, L. EL GHORMLI, and N. ROTHMAN. ‘Assessing the Probability that a Positive Report is False: An Approach for Molecular Epidemiology Studies.’ Journal of the National Cancer Institute 96 (2004): 434-42.

WILDE, L. ‘On the Use of Laboratory Experiments in Economics.’ The Philosophy of Economics, 1980.

ZAKOLYUKINA, A. ‘Measuring Intentional GAAP Violations: A Structural Approach.’ Working paper, 2015.