NBER WORKING PAPER SERIES

ON THE GENERALIZABILITY OF EXPERIMENTAL RESULTS IN ECONOMICS

Omar Al-Ubaydli
John A. List

Working Paper 17957
http://www.nber.org/papers/w17957

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2012

This paper is written for Frechette, G. & Schotter, A., Methods of Modern Experimental Economics, Oxford University Press. We wish to thank Marco Castillo, Robert Chambers, David Eil and Andreas Ortmann for helpful comments and for encouraging us to work on this issue. Alec Brandon and David Novgorodsky provided excellent research assistance. Al-Ubaydli: Department of Economics and Mercatus Center, George Mason University; List: Department of Economics, University of Chicago & NBER. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

On the Generalizability of Experimental Results in Economics
Omar Al-Ubaydli and John A. List
NBER Working Paper No. 17957
March 2012
JEL No. C9,C91,C92,C93,D03

ABSTRACT

Economists are increasingly turning to the experimental method as a means to estimate causal effects. By using randomization to identify key treatment effects, theories previously viewed as untestable are now scrutinized, efficacy of public policies are now more easily verified, and stakeholders can swiftly add empirical evidence to aid their decision-making. This study provides an overview of experimental methods in economics, with a special focus on developing an economic theory of generalizability. Given that field experiments are in their infancy, our secondary focus pertains to a discussion of the various parameters that they identify, and how they add to scientific knowledge. We conclude that until we conduct more field experiments that build a bridge between the lab and the naturally-occurring settings of interest we cannot begin to make strong conclusions empirically on the crucial question of generalizability from the lab to the field.

Omar Al-Ubaydli
Department of Economics and Mercatus Center
George Mason University
omar@omar.ec

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

On the generalizability of experimental results in economics

Omar Al-Ubaydli and John A. List[1]

March 2012

Abstract

Economists are increasingly turning to the experimental method as a means to estimate causal effects. By using randomization to identify key treatment effects, theories previously viewed as untestable are now scrutinized, efficacy of public policies are now more easily verified, and stakeholders can swiftly add empirical evidence to aid their decision-making. This study provides an overview of experimental methods in economics, with a special focus on developing an economic theory of generalizability. Given that field experiments are in their infancy, our secondary focus pertains to a discussion of the various parameters that they identify, and how they add to scientific knowledge. We conclude that until we conduct more field experiments that build a bridge between the lab and the naturally-occurring settings of interest we cannot begin to make strong conclusions empirically on the crucial question of generalizability from the lab to the field.

JEL codes: C90, C91, C93

Keywords: lab and field experiments; generalizability

1. Introduction

> The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past (Knight 1921, p313).

More than fifteen years ago one of the coauthors (List) sat in the audience of a professional presentation that was detailing whether and to what extent students collude in the lab and what this meant to policymakers interested in combating collusion. He openly wondered how such behavior would manifest itself with live traders in an extra-lab market, asking innocently whether policymakers should be concerned that this environment was much different than the one in which they typically operate. His concerns were swept aside as naïve.

[1] This paper is written for Frechette, G. & Schotter, A., *Methods of Modern Experimental Economics*, Oxford University Press. We wish to thank Marco Castillo, Robert Chambers, David Eil and Andreas Ortmann for helpful comments and for encouraging us to work on this issue. Alec Brandon and David Novgorodsky provided excellent research assistance. Al-Ubaydli: Department of Economics and Mercatus Center, George Mason University; List: Department of Economics, University of Chicago & NBER.

Later in that same year List attended a conference where experimental economists debated the merits of an experimental study that measured the magnitude of social preferences of students. He asked if such preferences would thrive in naturally-occurring settings, and how they would affect equilibrium prices and quantities. In not so many words, he was told to go and sit in the corner again. After the session, another junior experimentalist approached a now distraught List—"those are great questions, but off limits." List queried why, to which he received a response "that's the way it is."[2]

Except for the names and a few other changes, List was articulating words in the spirit of what Knight had much more eloquently quipped over 80 years prior: the intriguing possibility of using laboratory experiments as a solution to real world problems depended on the lab being like the field in terms of delivering similar behavioral relationships. A wet behind the ears List was fascinated by this query, but was learning that others did not share his passion, or even his opinion that it was a worthwhile point to discuss.

We are happy to find that the good ol' days are behind us. Today it is not uncommon for the very best minds in economics to discuss and debate the merits of the experimental method and the generalizability of experimental results (e.g., Falk and Heckman 2009, and the excellent chapters in Frechette and Schotter, forthcoming). We find this fruitful for many reasons, and continue to scratch our heads when some critics continue to contend that we have 'ruined the field of experimental economics' by scribing the original Levitt and List (2007; henceforth LL) article. This is a very short run view; indeed, our field of experimental economics can be sustainable only if our audience includes those outside our direct area of study. Otherwise, we run the real risk of becoming obscure. Understanding the applicability of our empirical results and having an open discussion can move us closer to the acceptance of our tools by all economists, and can move us toward an approach that can help us more fully understand the economic science.

More broadly, the discussions in Frechette and Schotter (forthcoming) represent a sign of change—we have entered a climate of scientific exploration that permits a serious investigation of what we believe to be the most important questions facing behavioral and experimental economists: (1) which insights from the lab generalize to the extra-lab world? (2) how do market interactions or market experience affect behaviors? And, (3) do individual anomalous behaviors aggregate to importantly affect market equilibria, and how does equilibration affect the individual anomalies?

One of LL's contributions was to present a theoretical framework and gather empirical evidence that questioned the level, or point, estimates delivered by laboratory experiments in economics. As a point of discussion, they focused on the work within the area of the measurement of social preferences. LL's overarching points included arguments that the laboratory is especially well equipped to deliver qualitative treatment effects, or comparative static insights, but not well suited to deliver deep structural parameters, or precise point estimates. This is because such estimates critically depend on the properties of the situation, as they detailed with examples from economics and psychology experiments. In the end, LL argue that lab and field experiments are complements with each serving an important role in the discovery process (consistent with what List has argued in all of his work).

[2] Without any evidence, we suspect that Peter Bohm was feeling similar ostracism as he presented his (seminal) challenges to laboratory experimentalists in Europe without much traction.

In this study we begin by providing an overview of experimental methods in economics, focusing on the behavioral parameters that each estimates. We then turn to formalizing generalizability. In principle, generalizability requires no less of a leap of faith in conventional (non-experimental) empirical research than in experimental research. The issue is obfuscated in non-experimental research by the more pressing problem of identification: how to correctly estimate treatment effects in the absence of randomization.

In our model, we generalize the 'all causes' approach to a more continuous form where researchers have priors about causal effects and update them based on data. This formality is necessary for a precise articulation of a theory of the advantages offered by field experiments. We conclude with some thoughts on where we hope this line of research goes in the coming years.

2. Preamble: Empirical methods

The empirical gold standard in the social sciences is to estimate a causal effect of some action. For example, measuring the effect of a new government program or answering how a new innovation changes the profit margin of a firm are queries for the scientist interested in causal relationships. The difficulty that arises in establishing causality is that either the action is taken or it is not—we never directly observe what would have happened in an alternative state in which a different action is taken. This, combined with the fact that in the real world there are simultaneously many moving parts, has led scholars to conclude that experimentation has little hope within economics.

Such thoughts reflect a lack of understanding of how the experimental method identifies, and measures, treatment effects. In fact, complications that are difficult to understand or control represent key reasons *to conduct* experiments, not a point of skepticism. This is because randomization acts as an instrumental variable, balancing unobservables across control and treatment groups.

To show this point, we find it instructive to consider empirical methods more broadly. The Easternmost portion of Figure 1, which we have often used elsewhere, highlights some of the more popular approaches that economists use to analyze naturally-occurring data.

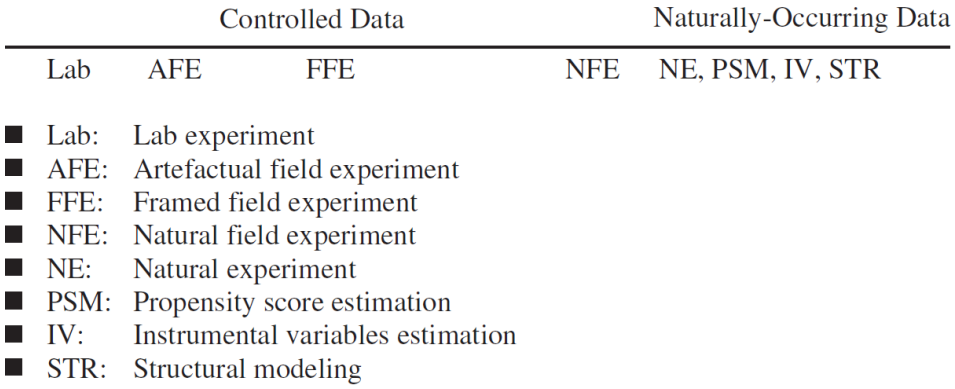

Figure 1: A field experiment bridge

For example, identification in natural experiments results from a difference-in-difference (DD) regression model where the major identifying assumption is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with treatment status, and that selection into treatment is independent of the temporary individual-specific effect. For example, let's say that the researcher is interested in estimating the impact on labor supplied from an increase in minimum wage, as Card and Krueger (1994) famously do by comparing labor supplied at fast food restaurants in New Jersey—which raised their minimum wage—and neighboring Pennsylvania—which did not change their minimum wage. There's no ex ante reason to expect New Jersey and Pennsylvania to start with the same labor supplied, but the motivation behind using DD is that you would expect the difference in labor supplied from year to year in both states to be pretty similar, all else equal.

Card and Krueger leverage the policy change in New Jersey to compare the difference of those differences in order to understand the impact of minimum wage laws on the quantity of labor supplied. Implicit in their analysis, though, is that other than the change in minimum wage laws in New Jersey, nothing has impacted the difference in the quantity of labor supplied between the time periods in Pennsylvania that is correlated with treatment. Furthermore, they must assume that treatment was randomly applied to New Jersey and not Pennsylvania, otherwise we don't know whether New Jersey just has some unique trait that is correlated with treatment status that would impact the quantity of labor supplied.

Useful alternatives to this approach include the method of propensity score matching (PSM) developed in Rosenbaum and Rubin (1983). A major assumption under this approach is called the "conditional independence assumption," and intuitively means that selection into treatment occurs only on observables. This means, for example, that the econometrician knows all the variables that influence whether a person selects into an employment program. In most cases, this assumption is unrealistic. Other popular methods of measurement include the use of instrumental variables and structural modeling. Assumptions of these approaches are well documented and are not discussed further here (see, e.g., Rosenzweig and Wolpin 2000 and Blundell and Costa Dias 2002).

We think that it is fair to say that these approaches of modeling naturally-occurring data are very useful, but because the world is complicated they are sometimes subject to incredulous assumptions. We are not the first to make this point, as there are entire literatures discussing the limitations of the various empirical models. In essence, many people argue that because the economic world is extremely complicated, one must take great care when making causal inference from naturally-occurring data.

On the Westernmost portion of Figure 1 is the laboratory experiment, which typically makes use of randomization to identify a treatment effect of interest among student subjects. Making generalizations outside of this domain might prove difficult in some cases, but to obtain the effect of treatment in this particular domain the only assumption necessary is appropriate randomization.

Field experiments represent a movement to take the data generation process beyond the walls of the laboratory. Two decades ago, the primary data generators were lab experimentalists. The past 15 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. Such subjects are non-standard in the sense that they are not students, but participants drawn from the market of interest. This type of experiment represents a useful type of exploration beyond traditional laboratory studies. As discussed in Frechette and Schotter (forthcoming), AFEs have been fruitfully used in financial applications, public economics, environmental economics, industrial organization, and to test predictions of game theory.

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. This type of experiment is important in the sense that a myriad of 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 one by one.

FFEs represent a very active type of field experiment in the past decade. Social experiments and recent experiments conducted in development economics are a type of FFE: subjects are aware that they are taking part in an experiment, and in many cases understand that their experience is for research purposes. Peter Bohm was an early experimenter to depart from traditional lab methods by using FFEs (Bohm 1972). While his work touched off an interesting stream of research within environmental and resource economics, for a reason that we cannot quite put our finger on, the broader economics literature did not quickly follow Bohm's lead to pursue research outside of the lab. This has only happened in the past decade or so.

Finally, a natural field experiment (NFE) is the same as a FFE in that it occurs in the environment where the subjects naturally undertake these tasks, but where the subjects *do not know* that they are participants in an experiment.[3] Such an exercise is important in that it represents an approach that combines the most attractive elements of the experimental method and naturally-occurring data: randomization and realism. In addition, it importantly tackles a selection problem that is not often discussed concerning the other types of experiments, as discussed below.

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). The economics of charity has witnessed a plethora of NFEs, as recently discussed in List (2011a). Of course, the taxonomy in Figure 1 leaves gaps, and certain studies may not fall neatly into such a classification scheme, but such an organization highlights what is necessary in terms of scientific discovery to link controlled experimentation to naturally-occurring data.

As we will argue below, a NFE represents the cleanest possible manner in which to estimate the treatment effect of interest. In this light, economists can certainly go beyond activities of astronomers and meteorologists and approach the testing of laws akin to chemists and biologists. Importantly, however, background variables can matter greatly when one attempts to generalize empirical results. With an

[3] This raises the issue of informed consent. For a discussion on this, and related, issues see Levitt and List (2009) and List (2008, 2011b).

understanding of the exact behavioral parameters identified by the various experimental approaches, we will be in a position to discuss generalizability, the focus of this paper. We first turn to the estimated parameters from experiments.

What parameters do experiments estimate?

Without loss of generality, 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. 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 states of the world. We assume that $p = 1$ indicates participation in the experiment, $p = 0$ indicates non-participation. That is, people who agree to enroll in the experiment have $p = 1$, others have $p = 0$. In this way, if one is interested in the mean differences in outcomes, then the treatment effect of interest is given by:

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

Yet, in our experience in the field, what is typically reported by government programs such as *Head Start*, firms—non-profits and for profits—and laypeople who discuss results from experiments is a treatment effect as follows:

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

Such a reported effect represents a potentially misleading measurement 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 $E(y_0|\, p = 1)$ from t', yielding:

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

where δ is the traditional selection bias term. This bias is a result of the non-treated differing from one another 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. For example, if parents who care more deeply about their children's educational outcomes are those who are more likely to sign up for services from *Head Start*, then their children might have better outcomes in the non-treatment state than children of parents who care less deeply about their children's educational outcomes. In this case, such selection bias causes the second term to be greater than zero because $E(y_0|\, p = 1) > E(y_0|\, p = 0)$, leading the *Head Start* program to report a treatment effect that is too optimistic; or a treatment effect estimate that is biased upwards. In such instances, we would systematically believe that the benefits of *Head Start* are considerably higher than their true benefits. 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, what is necessary is for randomization and identification of the treatment effect to occur just over the $p = 1$ group, yielding a treatment effect estimate of the mean

outcome differences between treated and non-treated from the $p = 1$ group. Letting $D = 1\ (0)$ denote those randomized into treatment (non-treatment):

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

At this point, it is instructive to pause and ask how to interpret the meaning of this treatment effect. First, this is the treatment effect that laboratory experiments, as well as AFEs and FFEs report (but not the treatment effect reported from NFEs). 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 made.

For example, 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 treatment status and correlated with the outcome variable, such generalization is frustrated. In our *Head Start* example, it might be the case that parents who believe *Head Start* will have a positive effect on their child are more likely to enroll. In that case, it would not be appropriate to generalize the effect from the $p = 1$ group to the $p = 0$ group if such beliefs were actually true.

This effect—call it 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, this sort of bias in the measured treatment effect is related to outcomes of the $p = 1$ and $p = 0$ groups in the *treated* state.

So how do NFEs differ in their identification approach? Since subjects are not aware 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 it. This fact excludes the typical selection effect discussed in the literature and precludes Treatment Specific Selection Bias (see Slonim et al. 2012 for a recent excellent study of selection into the laboratory). 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.

The very nature of how the 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 a myriad or one. 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. Indeed, all it takes is for one unobservable to be correlated with the outcome for an approach to have a problem of inference. The beauty behind randomization is that it handles the unobservability problem, permitting a crisp estimate of the causal effect of interest.

3. Formalizing generalizability

When we first began to explore generalizability, we found a dearth of theory and smattering of empirical evidence.[4] Even though we presented a theoretical framework in LL, our attention there was focused on the empirical evidence. Accordingly, here we focus on the theory and leave it to the interested reader to scrutinize the extant literature and make an informed opinion about what it says. Our own opinion is that it is too early to tell decisively where the empirical debate will end, but the evidence is mounting in favor of the hypotheses in LL. But, as usual, caveat lector—we leave it to the reader to decide.

In the all causes model (Heckman 2000), the researcher starts with a causal effect about which she has no prior. The purpose of an empirical investigation is to generate an estimate. In this section, we will generalize the all causes model to a more continuous form where researchers have priors about causal effects and update them based on data. This formality is necessary for a precise articulation of a theory of the advantages offered by field experiments; it is also consonant with our empirical complement presented below.

Setup

Let Y be a random variable, denoted the **dependent variable,** whose realizations are in $S_Y \subseteq \mathbb{R}$; let X be a random variable, denoted the **explanatory variable of interest,** whose realizations are in $S_X \subseteq \mathbb{R}$; and let Z be a random vector, denoted the **additional explanatory variables**, whose realizations are in $S_Z \subseteq \mathbb{R}^k$. Further, Z contains all the explanatory variables (apart from X) that have an impact on Y. To focus our model on the generalizability problem (rather than the sampling/inference problem), we assume that Z is observable. This model can be easily expanded to allow for unobservable variables.

In the all causes model, (X, Y, Z) are related according to the function $f: S_X \times S_Z \to S_Y$. Each $(x, x', z) \in S_X \times S_X \times S_Z$ is denoted a **causal triple**. The **causal effect** of changing X from x to x' on Y given $Z = z$ is described by the function $g: S_X \times S_X \times S_Z \to \mathbb{R}$, where:

$$g(x, x', z) = f(x', z) - f(x, z)$$

Let $T \subseteq S_X \times S_X \times S_Z$ be the **target space**. It describes the causal triples in which an empirical researcher is interested. Typically, she wants to know the exact value of the causal effect, $g(x, x', z)$, of each element of T. Often, particularly in experimental research, a researcher is interested merely in knowing if the causal effect lies in a certain range. Let $h: S_X \times S_X \times S_Z \to \mathbb{R}$ be a function that captures the aspect of a causal effect in which the researcher is interested. The most common, especially when testing theory (rather than selecting policy), is $\bar{h}$:

[4] Various people use the term external validity. As we noted in Harrison and List (2004, p1033), we do not like the expression "external validity" because "what is valid in an experiment depends on the theoretical framework that is being used to draw inferences from the observed behavior in the experiment. If we have a theory that (implicitly) says that hair color does not affect behavior, then any experiment that ignores hair color is valid from the perspective of that theory. But one cannot identify what factors make an experiment valid without some priors from a theoretical framework, which is crossing into the turf of "internal validity." Note also that the "theory" we have in mind here should include the assumptions required to undertake statistical inference with the experimental data."

$$\bar{h}(x,x',z)=\begin{cases}-1 \text{ if } g(x,x',z)<0\\ 0 \text{ if } g(x,x',z)=0\\ 1 \text{ if } g(x,x',z)>0\end{cases}$$

Before embarking upon a new empirical investigation, a researcher has a **prior** $F^0_{x,x',z}:\mathbb{R}\to[0,1]$ about the value of $h(x,x',z)$ for each $(x,x',z)\in T$. The prior is a cumulative density function based on existing theoretical and empirical studies, as well as researcher introspection.

An empirical investigation is a **dataset** $D\subseteq S_X\times S_X\times S_Z$. Note that D and T may be disjoint, and both may be singletons. Indeed, D is often a singleton in laboratory experiments. The researcher will typically sample Y repeatedly at $(X,Z)=(x,z)$ and $(X,Z)=(x',z)$ and use this to obtain an estimate of $g(x,x',z)$. Let the **results** $R\subseteq D\times\mathbb{R}$ be the set of causal effects obtainable from the dataset D *making no parametric assumptions* (i.e., no extrapolation or interpolation):

$$R=\{(x,x',z,g(x,x',z)):(x,x',z)\in D\}$$

As mentioned above, we set aside the sizeable problem of obtaining a consistent estimate of $g(x,x',z)$. In fact this is the primary problem faced by most non-experimental, empirical research due to, e.g., small samples and endogeneity problems. To some extent, generalizability is a secondary issue in empirical research that uses naturally-occurring data simply because it is overshadowed by the more pressing issue of identification.

This essay will ignore this part of the identification problem to focus attention upon the generalizability problem. Questions about how sample size and variance affect the estimation procedure are set aside as they do not interact with the main principles, though this framework can be easily expanded to incorporate such issues. Consequently, we do not draw a distinction between a causal effect $g(x,x',z)$ and a direct empirical estimate of $g(x,x',z)$.

After seeing the results, R, the researcher updates her prior $F^0_{x,x',z}$ for each $(x,x',z)\in T$, forming a **posterior** $F^1_{x,x',z}$. *The updating process is not necessarily Bayesian*. The generalizability debate, which we discuss in the next section, is concerned with the formation of the posterior, especially for elements of $T\backslash D$. We henceforth assume that the prior is never completely concentrated at the truth, implying that any valid estimate of $g(x,x',z)$ will always lead to the researcher updating her prior.

The posterior is the conclusion of the empirical investigation. This framework is designed to include studies that estimate causal effects for policy use, for testing a theory or for comparing multiple theories.

To put the framework into motion with an economic example, we consider a Laffer-motivated researcher who wants to know if increasing sales tax (X) from 10% to 15% increases tax revenue (Y) when the mean income in a city (Z) is \$30k. For expositional simplicity, we assume that the only element of Z is mean income level. The researcher can only generate data in four cities: two cities have a mean income of \$20k and two cities have a mean income of \$35k. All four cities currently have a sales tax of 10%. She randomly assigns treatment (increasing sales tax to 15%) to one city in each income pair and control (leaving the sales tax at 10%) to the other city in each pair. She then collects data on tax revenue (one observation in each cell is sufficient because we are not tackling the sample-size component of the identification problem).

The researcher's prior is a 0.5 chance of a positive causal effect at a mean income of \$30k. She finds a positive causal effect at both mean income levels and revises her prior at a mean income of \$30k to a 0.6 chance of a positive causal effect. In terms of our notation:

$$T = \{(10\%, 15\%, \$30000)\}$$

$$h(x, x', z) = \begin{cases} 1 \text{ if } g(x, x', z) > 0 \\ 0 \text{ if } g(x, x', z) \leq 0 \end{cases}$$

$$D = \{(10\%, 15\%, \$20000), (10\%, 15\%, \$35000)\}$$

$$R = \{(10\%, 15\%, \$20000, 1), (10\%, 15\%, \$35000, 1)\}$$

$$F^0_{10\%,15\%,\$30000}(0) = 0.5, F^1_{10\%,15\%,\$30000}(0) = 0.4$$

Different types of generalizability

Given a set of priors $\mathcal{F}^0 = \{F^0_{x,x',z} : (x, x', z) \in S_X \times S_X \times S_Z\}$ and results R, the **generalizability set** $\Delta(R) \subseteq \{S_X \times S_X \times S_Z\} \backslash D$ is the set of causal triples outside the dataset where the posterior $F^1_{x,x',z}$ is updated as a consequence of learning the results:

$$\Delta(R) = \{(x, x', z) \in \{S_X \times S_X \times S_Z\} \backslash D : F^1_{x,x',z}(\theta) \neq F^0_{x,x',z}(\theta) \text{ for some } \theta \in \mathbb{R}\}$$

Results are **generalizable** when the generalizability set is non-empty: $\Delta(R) \neq \emptyset$. A researcher is said to **generalize** when the generalizability set intersects with the target space: $\Delta(R) \cap T \neq \emptyset$. The researcher in the above Laffer example is generalizing. Note that generalizability is focused on $h(x, x', z)$ rather than $g(x, x', z)$ since the prior is focused on $h(x, x', z)$.

As mentioned above, in principle, generalizability requires no less of a leap of faith in conventional (non-experimental) empirical research than in experimental research. The issue is obfuscated in non-experimental research by the more pressing problem of identification: how to correctly estimate $g(x, x', z)$ in the first place due to, e.g., the absence of randomization. This problem does not plague experimental work. Indeed, the beauty of experimentation is that through randomization the problem of identification is solved.

Given prior beliefs $\mathcal{F}^0$, a set of results R has **zero generalizability** if its generalizability set is empty: $\Delta(R) = \emptyset$. Zero generalizability is the most conservative empirical stance and equates to a paralyzing fear of interpolation, extrapolation, or the assumption of additive separability.

Given prior beliefs $\mathcal{F}^0$, a set of results R has **local generalizability** if its generalizability set contains points within an arbitrarily small neighborhood of points in D:

$$(x, x', z) \in \Delta(R) \Rightarrow (x, x', z) \in B_\varepsilon(\bar{x}, \bar{x}', \bar{z}) \text{ for some } \varepsilon > 0, (\bar{x}, \bar{x}', \bar{z}) \in D$$

The simplest way to obtain local generalizability is to assume that $h(x, x', z)$ is continuous (or only has a small number of discontinuities), since continuity implies local linearity and therefore permits local

extrapolation.[5] In the Laffer example above, assuming that the causal effect is continuous in the mean income level in the city, the researcher can extrapolate her findings to estimate the causal effect for a city with a mean income level of $35100. In principle, non-local changes in (x, x', z) can have a large effect on h, limiting our ability to extrapolate. However *as long as we do not change* (x, x', z) *by much and* $h(x, x', z)$ *is continuous, then* h *will not change by much* and so our dataset D will still be informative about causal effects outside this set.

Since continuity is sufficient for local generalizability, it follows that discontinuity is necessary for zero generalizability. If, as is often likely to be the case, the researcher is unsure of the continuity within $h(x, x', z)$, then the more conservative she is, the more she will be inclined to expect zero generalizability.[6]

Given prior beliefs $\mathcal{F}^0$, a set of results R has **global generalizability** if its generalizability set contains points outside an arbitrarily small neighborhood of points in D:

$$\exists (x, x', z) \in \Delta(R): (x, x', z) \notin B_\varepsilon(\bar{x}, \bar{x}', \bar{z}) \text{ for some } \varepsilon > 0, \text{for all } (\bar{x}, \bar{x}', \bar{z}) \in D$$

In the Laffer example above, the researcher is assuming global generalizability. At its heart, *global generalizability is about assuming that a large change in* (x, x', z) *does not have a large effect on* h.

A succinct summary of Section 3 thus far is as follows.

1. In a non-parametric world, results can fail to generalize, generalize locally, or generalize globally.
2. A sufficient condition for local generalizability is continuity of $h(x, x', z)$.
3. A sufficiently conservative researcher is unlikely to believe that her results generalize globally because this requires a much stronger assumption than continuity.

We are now in a position to formalize the advantages offered by field experiments.

A theory of the advantage offered by field experiments

A (function of a) causal effect $h(x, x', z)$ is **investigation-neutral** if it is unaffected by the fact that it is being induced by a scientific investigator *ceteris paribus*. Thus, for example, suppose that we are studying the causal effect of the slope of a demand curve on the percentage of surplus realized in a market. If this effect is investigation-neutral, then the fact that the market was set up as the result of a scientific investigation versus simply observed in the naturally-occurring domain, ceteris paribus, does not change the causal effect. **We assume that causal effects are investigation-neutral**.

We define a **natural setting** as a triple (x, x', z) that can plausibly exist in the absence of academic, scientific investigation. For example if a scientist is studying the effect of a piece rate versus a fixed wage compensation scheme on the productivity of a worker soliciting funds in a phoneathon for a charity, then this is a natural setting since it is common for workers to get hired to do such tasks using a piece rate or a fixed wage scheme. In contrast, if a scientist is interested in studying the magnitude of social preferences

[5] Continuity in a subset of its arguments guarantees local generalizability in a subset of dimensions.
[6] This is where our allowance for non-Bayesian updating applies; a highly conservative researcher may be reluctant to update her prior if there is a large probability of the generalization being invalid.

and brings a group of students into the lab to play a dictator game, then this is not a natural setting since students virtually never find themselves involved in such a scenario under the specific features of that environment and task.

Our principal assumption is that as economists, we are more interested in learning about and understanding behavior in natural settings than in non-natural settings. This does not eliminate the value of learning about causal effects in non-natural settings; after all, the benefits of centuries of artificial studies in physics, chemistry, and engineering are self-evident. However it requires that insights gained in non-natural settings generalize to natural settings for them to be of great value. This is because as economists we are interested with reality, in contrast to say poetry. We are concerned with understanding the real world and in modifying it to better the allocation of scarce resources or to prescribe better solutions to collective choice problems.

Through this lens, because of their very nature laboratory experiments represent an environment that could only ever come about as the result of a scientific investigation. Thus, **laboratory investigations are not completed in natural settings**. Moreover, many laboratory experiments might not **even be in the neighborhood of a natural setting**. This is because several variables have to change by large amounts in order for a laboratory setting to transform into a natural setting, e.g., the nature and extent of scrutiny, the context of the choice decision and situation, the experience of participants, and several other factors discussed in LL. We elaborate on one such factor—the participation decision—below.

Falk and Heckman (2009) and others (see the work in Frechette and Schotter, forthcoming) have questioned whether the non-local changes in (x, x', z) that arise when generalizing from a laboratory setting to field setting have a large effect on $h(x, x', z)$. Interestingly, when making their arguments they ignore one of the most important: typical laboratory experiments impose artificial restrictions on choice sets and time horizons.

Regardless of the factors that they discuss and fail to discuss, to the best of our knowledge, nobody has questioned the proposition that the changes in (x, x', z) are non-local.[7] In fact, the artificial restrictions on choice sets and time horizons are a particularly dramatic illustration of the non-local differences between laboratory and field settings. Another critical, non-local difference between laboratory and natural field settings is the participation decision, shown above in the traditional treatment effects model and discussed below within our framework.

With this background in hand, we proceed to three Propositions which are meant to capture the range of thoughts across the economics profession today. We do not believe that one can categorize all laboratory experiments under any one of these propositions, but rather believe that there are a range of laboratory experiments, some of which fall under each of the three propositions.

Proposition 1: Under a liberal stance (global generalizability), neither field nor laboratory experiments are demonstrably superior to the other.

[7] We are therefore implicitly referring to NFEs (Harrison and List 2004) when we discuss field experiments in this section, since FFEs and AFEs are not natural settings in every dimension. However in Propositions 1-3, they will lie between NFEs and conventional laboratory experiments.

This view is the most optimistic for generalizing results from the lab to the field. It has as its roots the fact that the generalizability sets are both non-empty and, in general, neither will contain the other. In this way, empirical results are globally generalizable.

As an example, consider the work on market equilibration. Conventional economic theory relies on two assumptions: utility-maximizing behavior and the institution of Walrasian tâtonnement. Explorations to relax institutional constraints have taken a variety of paths, with traditional economic tools having limited empirical success partly due to the multiple simultaneously moving parts in the marketplace. Vernon Smith (1962) advanced the exploration significantly when he tested neoclassical theory by executing double oral auctions. His results were staggering—quantity and price levels were very near competitive levels after a few market periods. It is fair to say that this general result remains one of the most robust findings in experimental economics today.

List (2004) represents a field experiment that moves the analysis from the laboratory environment to the natural setting where the actors actually undertake decisions. The study therefore represents an empirical test in an actual marketplace where agents engage in face-to-face continuous bilateral bargaining in a multi-lateral market context.[8] Much like Smith's (1962) set-up, the market mechanics in List's bilateral bargaining markets are not Walrasian.

Unlike Smith (1962), however, in these markets subjects set prices as they please, with no guidance from a centralized auctioneer. Thus, List's design shifts the task of adaptation from the auctioneer to the agents, permitting trades to occur in a decentralized manner, similar to how trades are consummated in actual free unobstructed markets. In doing so, the market structure reformulates the problem of stability of equilibria as a question about the behavior of actual people as a psychological question—as opposed to a question about an abstract and impersonal market.

A key result of List's study is the strong tendency for exchange prices to approach the neoclassical competitive model predictions, especially in symmetric markets. This example highlights exactly what the original LL model predicts: a wide class of laboratory results should be directly applicable to the field. In particular, we would more likely find an experiment falling under Proposition 1 when the experimenter does not place the subject on an artificial margin, when moral concerns are absent, the computational demands on participants are small, non-random selection of participants is not an important factor, experience is unimportant or quickly learned, and the experimenter has created a lab context that mirrors the important aspects of the real-world problem. At that point, we would expect results from the lab to be a closer guide to natural settings.

Our next Proposition strengthens this liberal view:

Proposition 2: Under a conservative stance (local generalizability; or if the researcher is confident that $h(x, x', z)$ is continuous), field experiments are more useful than laboratory experiments.

This view follows from the idea that results generalizable locally. Thus, whether empirical data is generated in the lab or the field, it can be generalized to the immediately adjacent settings. And, since

[8] In this way, List's (2004) institution was more in line with Chamberlin (1948) than Smith. Since Chamberlin's original lab results have proven not to replicate well, we view his laboratory insights as an aberration when discussing lab results from market experiments.

field experiments provide information from a natural setting and laboratory experiments from a non-natural setting, field experiment are more useful. This is because the neighborhood of a natural setting is still a natural setting, while the neighborhood of a non-natural setting is non-natural.

As an example, consider the recent work in the economics of charity. Without a doubt, the sector represents one of the most vibrant in modern economies. In the US alone, charitable gifts of money have exceeded 2% GDP in the past decade. Growth has also been spectacular—from 1968-2008, individual gifts have grown nearly 18 fold, doubling the growth rate in the S&P 500. Recently, a set of lab and field experiments have lent insights into the "demand side" of charitable fundraising.

For instance, consider the recent laboratory experiments of Rondeau and List (2008). They explored whether leadership gifts—whether used as a challenge gift (simply an announcement) or as a match gift (i.e., send in $100 and we will double your contribution)—affect giving rates. From the lab evidence, they found little support for the view that leadership gifts increase the amount of funds raised.

Alternatively, in that same paper, they used leadership gifts to raise money for the Sierra Club of Canada via a field experiment. Their natural field experiment was conducted within the spirit of one of the typical fundraising drives of the Sierra Club organization. A total of 3,000 Sierra Club supporters were randomly divided into four treatments, varying the magnitude and type of leadership gift. They find that challenge gifts work quite well in the field. This means that it is important for fundraisers to seek out big donors privately before they go public with their cause, and to use challenge gifts when doing so.

One is now in a position to ask: if I am a fundraiser, which set of results should guide my decision-making—those from the lab or the field?

Viewed through the lens of Proposition 2, practitioners in the field who are interested in raising money for their cause would be well served to pay close attention to the field experimental results because such insights are locally generalizable (see also List 2011a). On the other hand, the lab results that suggest the upfront monies raised will not help much *are less likely* to generalize outside of the lab confines.

This result highlights that economists are often only concerned with obtaining the sign of a causal effect $g(x, x', z)$, as summarized by the function $\bar{h}(x, x', z)$ above. In this case, if the researcher is confident that $g(x, x', z)$ is monotonic in z_i over some range $[z_{i0}, z_{i1}]$, then $\bar{h}(x, x', z)$ will be continuous almost everywhere. This is sufficient for local generalizability.

Finally, an even further tightening of the restriction set leads to our third Proposition:

Proposition 3: Under the most conservative stance (zero generalizability), field experiments are more useful than laboratory experiments because they are performed in one natural setting.

This cautious view has as its roots in the fact that nothing is generalizable beyond the specific context where the investigation occurs.[9] Thus, because field experiments are guaranteed to help us to refine our prior about one natural setting—the causal effect that the field experiment itself estimates—they are more

[9] Of course, an even more extreme view is to conclude that we can learn nothing from empirical work because of the passage of time.

useful. In contrast, under this level of conservatism, laboratory experiments tell us nothing about any natural setting.

Consider the increasingly-popular task of measuring social preferences. One popular tool to perform the task is a dictator game. The first dictator game experiment in economics is due to Kahneman, Knetsch, and Thaler (1986). They endowed subjects with a hypothetical $20, and allowed them to dictate either an even split of $20 ($10 each) with another student or an uneven split ($18, $2), favoring themselves. Only 1 in 4 students opted for the unequal split. Numerous subsequent dictator experimental studies with real stakes replicate these results, reporting that usually more than 60 percent of subjects pass a positive amount of money, with the mean transfer roughly 20 percent of the endowment.

The common interpretation of such findings can be found in Henrich et al.'s (2004) work: "Over the past decade, research in experimental economics has emphatically falsified the textbook representation of Homo economicus, with hundreds of experiments that have suggested that people care not only about their own material payoffs but also about such things as fairness, equity, and reciprocity." Indeed, the point estimates of giving from these experiments have even been used to estimate theoretical models of social preferences (see, e.g., Fehr and Schmidt 1999).

Under the extreme view of Proposition 3, such insights have limited applicability because the properties of the situation are such that we only learn about one specific situation—giving in the lab. In short, our model informs us that putting subjects on an artificial margin in such a setting necessarily limits the ability to make direct inference about markets of interest.

As a point of comparison, consider a recent field measurement of social preferences from List (2006a). As discussed more fully below, one of the goals of this study was to measure the importance of reputation and social preferences in a naturally-occurring setting. To explore the importance of social preferences in the field, List (2006a) carries out gift exchange natural field experiments in which buyers make price offers to sellers, and in return sellers select the quality level of the good provided to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers.

The results from the AFEs in List (2006a) mirror the typical laboratory findings with other subject pools: strong evidence consistent with social preferences was observed through a positive price and quality relationship. List (2006a) reports that similarly constructed FFEs provide identical insights. Yet, when the environment is moved to the marketplace via a NFE, where dealers are unaware that their behavior is being recorded as part of an experiment, little statistical relationship between price and quality emerges.

Viewed through the lens of Proposition 3, this study provides three social preference estimates that are applicable to *only* the three specific environments in which they are measured. The first estimate uses actual traders from this market in a laboratory experiment. The second uses actual traders from this market in a setting that resembles the market that they have naturally selected to participate, but one in which they know that they are being scrutinized. The third observes actual traders in a market that they have naturally selected to participate, wherein they do not know that they are being observed for scientific purposes. As such, under the extreme view of Proposition 3, we have at least learned about one naturally-occurring setting from List's (2006a) data.

Our three propositions are summarized visually in Figure 2. Consider a causal triple where we vary two of the dimensions of . The space is divided into natural environments (above the dashed line) and non-natural environments (below the dashed line). One combination of is the field experiment and one is the laboratory experiment, each of which is depicted by a spot in the figure.

Under conservative generalizability (the inner, black circles), only the field experiment yields information about natural environments. As we become less conservative and the circles expand (to the outer, gray circles), both types of experiments yield potentially disjoint information about natural environments. Thus, they become complements in the production of knowledge.

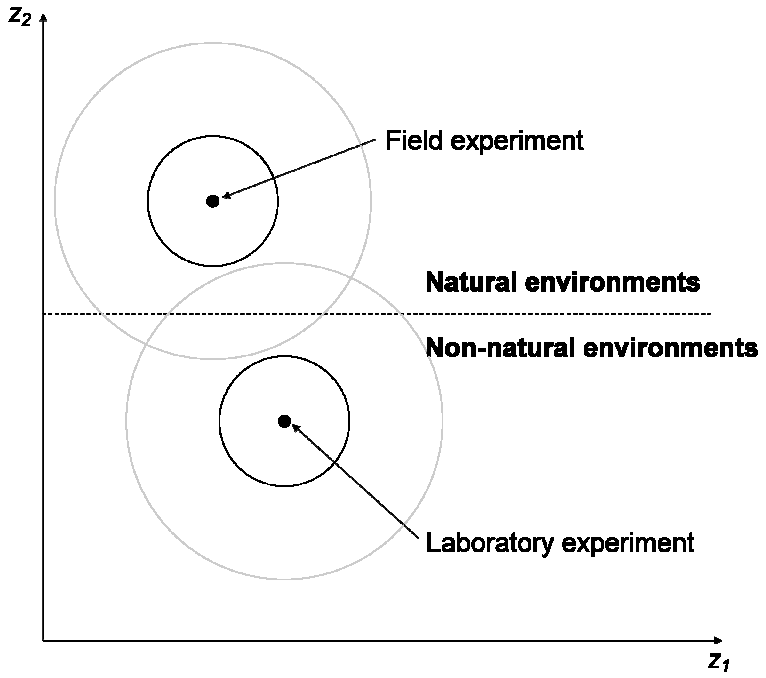

Figure 2: Generalizability in field and lab experiments

In the simpler version of the all-causes model, Falk and Heckman (2009) claim that generalizability requires an assumption of additive separability, an arbitrary assumption that is no more plausible for field experiments than it is for laboratory experiments. However their claim only applies for global generalizability; when generalizing locally under the assumption of continuity, additive separability is not necessary and the advantage of field experiments is particularly salient.

The kind of statistical conservatism required for zero- or local generalizability is extreme, and this is because we have a highly discontinuous definition of both: priors for certain subsets of have to be *completely* unchanged in response to non-intersecting data. A more realistic treatment would be to include a more continuous measure of generalizability. We used highly stylized, discontinuous measures purely for expositional simplicity, akin to summarizing a hypothesis test by its conclusion (accept or reject) rather than by the p-value associated with the test-statistic. The essence of our argument is unchanged by allowing generalizability to evolve into a more continuous concept.

Extending the model: The participation decision

In Section 2, we discussed how selection impacts the measurement of treatment effects. In this section, we return to this topic and use our formal structure to extend the previous treatment effects discussion on the participation decision.

Consider a family of causal triples $\{g(x, x', z)\}_{z \in U_Z \subseteq S_Z}$ that an investigator wants to estimate, where z is unidimensional. Z can be thought of as a potentially observable individual-level characteristic, such as preferences or IQ. In the absence of experimental interference by the investigator, individuals learn their realization of Z and can then influence the realization of X. For simplicity, assume that at a (potentially small) cost, they can guarantee the control value, $X = x$. We assume that it is the control rather than the treatment because usually, the treatment corresponds to an intervention, whereas the control is the status quo. Conditional on the realization of Z, all remaining randomness is exogenous. Assume that at every $z \in U_Z$, a positive proportion of people are observed in each of control and treatment: $\forall z \in U_Z, 0 < \Pr(X = x | Z = z) < 1$ and $0 < \Pr(X = x' | Z = z) < 1$.

At this point, in principle, no experiment need be conducted. Under our highly stylized framework, the investigator can simply collect two naturally-occurring observations at each value of Z (a control and a treatment) and thereby directly calculate $g(x, x', z)$. In practice, the investigator has to worry about sample sizes (the sampling issue that we abstracted away from above) and she may have a strict time limit for data collection, either of which would push her toward running an experiment where she directly and randomly manipulates the value of X.

If, after deciding to conduct an experiment, the investigator chooses to conduct it covertly (as in NFEs), then inference will proceed as normal and the desired family of causal effects will be estimated. Her ex post control over the value of X swamps individuals' ability to influence X.

On the other hand, should the investigator publicize her intention to conduct the experiment, then she has to worry about subjects exercising their ex ante control over X as a result of knowing about the experiment. Suppose some subset $U_Z' \subset U_Z$ decides to guarantee themselves the control value of X, meaning that the investigator cannot estimate the causal triples for this subset. The investigator has a large degree of control over X, but usually she cannot force those who, upon becoming aware of the experiment, choose not to participate. Inference for the remaining group, $U_Z \backslash U_Z'$, remains valid as before.

Consequently, she will be forced to update her priors on causal triples associated with U_Z' by extrapolating/interpolating from $U_Z \backslash U_Z'$. In practice, this will be rendered even more precarious by the possibility that Z is unobservable, meaning that the experimenter will be forced to assume that the causal triple is simply unaffected by the participation decision.[10] In the case when $U_Z = \{z_1, z_2\}, U_Z' = \{z_2\}$, the extrapolation bias, which we term Treatment Specific Selection Bias, will be:

$$B = g(x, x', z_2) - g(x, x', z_1)$$

[10] Of course, a selection model can limit the size of the necessary leap of faith. However unless the investigator can convincingly present a perfectly deterministic participation model, or one where residual randomness is definitively exogenous with respect to the treatment effect (neither of which is likely), then bias will remain a concern.

Thus ironically, in a specific sense, natural field experiments afford the investigator *more* control over the environment because it allows her to bypass the participation decision. *This insight is exactly opposite to received wisdom, wherein critics argue that field experiments have less control*.

This abstract argument is illustrated above with the *Head Start* example: if parents who care more deeply about their children's outcomes are more likely to sign up for services from *Head Start*, then their children might have better outcomes in the non-treatment state than children of parents who care less deeply about their children. This orthodox selection effect is what motivates the investigator to randomize. The investigator will publicize the randomized program and solicit for enrollment, creating the two groups $U_Z \backslash U_Z'$ (participants) and U_Z' non-participants. However it might be the case that parents who believe *Head Start* will have a significant effect on their child are more likely to enroll. In that case, it would not be appropriate to generalize the effect from the $U_Z \backslash U_Z'$ group to the U_z' group if such beliefs were actually true; the bias term B would be negative.

One potential example of this bias is randomization bias—where a direct aversion to the act of randomization is what discourages people from participating. This would be a valid concern for long-term studies where the ex ante uncertainty generated by randomization may lead to an expectation of adjustment costs and hence the certainty of non-participation is preferred.

More generally, due to cognitive limitations, people do not take too active a role in determining natural treatment allocation in many day-to-day decisions, and so there is room for covert experimentation, e.g., in how the goods are displayed in a grocery store or how a commercial looks on TV. But the very public declaration of a randomized control trial could signal the importance of a certain decision and motivate an individual to devote the cognitive resources necessary to exercise full control over participation. If you are convinced that the treatment of viewing a TV commercial is undesirable, you can just turn your TV off.

The covertness implicit in a NFE, which we are arguing is desirable, is sometimes impossible, especially in large, new programs where there is no natural, pre-existing target population whose natural choices over treatment and control can be subtly manipulated by an investigator. For example, if we wanted to estimate the causal effect of introducing neighborhood watch schemes in areas with few to no neighborhood watch schemes, participation is likely to be limited in a way that interacts with the treatment effect and in a way that cannot be circumvented by covertness.

Fortunately, it is possible in many fields of interest, such as design of incentive schemes across many important economic domains, charitable contributions, auction design, marketing, worker compensation, organizational structure, and so on.

Advantages of laboratory experiments

Despite Propositions 1-3, our model strongly shows that there is a critically important advantage of laboratory experiments over field experiments. Thus far, the target space T and dataset D are exogenous. As suggested in the previous section, in practice, many causal triples are inestimable in field settings due to ethical/feasibility/cost reasons. For example, it is straightforward to set up a model economy in the

laboratory and to manipulate randomly interest rates to gauge their effect on inflation. No such experiment is possible in a natural field experiment.

In this sense, the range of causal triples that cannot be directly estimated in a natural field experiment and that lie outside the local generalizability set of estimable causal triples is so large that in many environments, field and laboratory experiments become natural complements.[11]

Consider the case of discrimination. One would be hard-pressed to find an issue as divisive for a nation as race and civil rights. For their part, economists have produced two major theories for why discrimination exists: i) certain populations having a general "distaste" for minorities (Becker 1957) and ii) statistical discrimination (see, e.g., Arrow 1972, Phelps 1972), which is third-degree price discrimination as defined by Pigou: marketers using observable characteristics to make statistical inference about productivity or reservation values of market agents. Natural field experiments have been importantly used to measure and disentangle the sources of discrimination (see List 2006b for a survey).

Now consider how a laboratory experiment would be formulated. For example, if one were interested in exploring whether, and to what extent, race or gender influences the prices that buyers pay for used cars, it would be difficult to measure accurately the degree of discrimination among used car dealers who know that they are taking part in an experiment. We expect that in such cases most would agree that Propositions 2 or 3 hold.

This is not say that lab experiments cannot contribute to our understanding of important issues associated with discrimination. Quite the opposite. Consider the recent novel work of Niederle et al. (2008). They use lab experiments to investigate whether affirmative action changes the pool of entrants into a tournament. More specifically, they consider a quota system which requires that out of two winners of a tournament at least one be a woman. We suspect that this would be quite difficult to do legally in a natural field experiment. Interestingly, they report that the introduction of affirmative action results in substantial changes in the composition of entrants.

This is just one of many studies that we could point to that serves to illustrate that, once viewed through the lens of our model, laboratory and field experiments are more likely to serve as complements as most suspect.

An aspect of laboratory experimentation that is outside of our model and another important is the ease of replication. Since replication is the cornerstone of the experimental method, it is important to discuss briefly the power of replication. For the purposes of this exposition, suffice it to say that the greater ease of replication in the lab suggests an additional dimension of complementarity between field and lab experiments, particularly in the search for true qualitative results about causal relationships. We refer the interested reader to Maniadis, Tufano and List (2011) for a fuller discussion of replication and its benefits.

[11] Below we give an explicit example of an important case wherein a NFE estimates an effect that is difficult (perhaps impossible) to measure in the lab.

4. Epilogue

Going beyond parallelism and discussing scientifically the important issue of generalizability has been an invaluable turn for the better within experimental economics. Whereas empirical evidence is beginning to mount that helps to shed light on whether, and to what extent, received results generalize to other domains, there have been less theoretical advances. In this study, we put forth a theoretical model that helps frame the important features within the debate on generalizability. In doing so, it highlights the important role that field experiments should play in the discovery process.

Levitt and List (2009) discuss three distinct periods of field experiments in economics. The first period is encompassed predominantly by the work of Fisher and Neyman in the 1920s and 1930s. This period was seminal in that it helped to answer important economic questions regarding agricultural productivity while simultaneously laying the statistical groundwork relied on today. A second period of interest is the latter half of the 20th century, during which government agencies conducted a series of large-scale social experiments. In Europe, early social experiments included electricity pricing schemes in Great Britain in the late 60s. The first wave of such experiments in the U.S. began in earnest in the late 60s and included government agency attempts to evaluate programs by deliberate variations in policies. These experiments have had an important influence on policy, have generated much academic debate between structuralists and experimentalists, and anticipated the wave of recent field experiments executed in developing countries.

The third distinct period of field experimentation is the surge of field experiments in economics in the past decade or so. This most recent movement approaches field experiments by taking the tight controls of the lab to the field. Although in their infancy, the field experiments produced during this third period have already contributed to economic science by (1) measuring key parameters to test theory, and when the theory is rejected collected enough information to inform a new theory, (2) informed policymakers, (3) extended to both non-profit and for profit firms, and (4) being instrumental methodologically in bridging laboratory and non-experimental data. We believe going forward that field experiments will represent a strong growth industry as people begin to understand the behavioral parameters field experiments estimate and the questions they can address.

We believe that at this point social scientists can move beyond strong statements that *lab or field* results will always or never replicate. This type of reasoning seems akin to standing on the stern of the Titanic and saying she will never go down after the bow sinks below the water surface. Rather, it is now time to more fully articulate theories of generalizability and bring forward empirical evidence to test those theories. Building a bridge between the lab and the field is a good place to start. We hope that this paper and the discussion in Frechette and Schotter (forthcoming) move researchers to use AFEs, FFEs, and NFEs to bridge insights gained from the lab with those gained from modeling naturally-occurring data.

References

Arrow, K (1972). The Theory of Discrimination, in Discrimination in Labor Markets. O. Ashenfelter and A. Rees, eds., Princeton, NJ: Princeton University Press.

Becker, G.S. (1957). The economics of discrimination. (2nd ed.). Chicago, IL: University of Chicago Press.

Blundell, R., & Costa Dias, M. (2002). Alternative approaches to evaluation in empirical microeconomics. Portuguese Economic Journal, 1(2), 91-115.

Bohm, P. (1972). Estimating demand for public goods: An experiment. European Economic Review, 3(2), 111-130.

Card, D., & Krueger, A.B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review, 84(4), 772-793.

Chamberlin, E.H. (1948). An experimental imperfect market. Journal of Political Economy, 56(2), 95-108.

Falk, J. J. , & Heckman, A. (2009). Lab experiments are a major source of knowledge in the social sciences. Science, 326, 535-538.

Fehr, E., & Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3), 817-8.

Frechette, G. & Schotter, A. (Forthcoming). *Methods of Modern Experimental Economics*, Oxford University Press.

Harrison, G.W., & List, J.A. (2004). Field experiments. Journal of Economic Literature, 42(4), 1009-1055.

Heckman, J.J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. The Quarterly Journal of Economics, 115(1), 45-97.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford: Oxford University Press.

Hossain, T., & Morgan, J. (2006). ...Plus shipping and handling: Revenue (non) equivalence in field experiments on eBay. Advances in Economic Analysis & Policy, 6(2), Article 3.

Kahneman, D., Knetsch, J.L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. American Economic Review, 76(4), 728-741.

Knight, F. H. (1921). Risk, uncertainty, and profit. New York, NY: Cosimo.

Levitt, S.D., & List, J.A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? The Journal of Economic Perspectives, 21(2), 153-174.

Levitt, S.D., List, J.A. (2009). Field experiments in economics: The past, the present, and the future. European Economic Review, 53(1), 1-18.

List, J.A. (2003). Does market experience eliminate market anomalies? The Quarterly Journal of Economics, 118(1), 41-71.

List, J. A. (2004). Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. Econometrica, 72(2), 615-625.

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

List, J. A. (2006b). Field Experiments: A Bridge between Lab and Naturally Occurring Data. Advances in Economic Analysis & Policy, 6(2), Article 8.

List, J.A. (2008). Informed Consent in Social Science. Science, 322, 672.

List, J.A. (2011a). The Market For Charitable Giving. Journal of Economic Perspectives, 25(2), 157-180.

List, J.A. (2011b). Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off. Journal of Economic Perspectives, 25(3), 3-15**.**

Maniadis, Z., Tufano, F., & List, J.A. (2011). One Swallow Doesn't Make a Summer: How Economists (Mis-)Use Experimental Methods and Their Results. Working Paper.

Niederle, M., Segal, C., & Vesterlund, L. (2008). How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness. NBER working paper series (no. 13923).

Phelps, E.S. (1972). The statistical theory of racism and sexism. The American Economic Review, 62(4), 659-661.

Rondeau, D., & List, J.A. (2008). Matching and challenge gifts to charity: Evidence from laboratory and natural field experiments. Experimental Economics, 11, 253-267.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Rosenzweig, M.R., & Wolpin, K.I. (2000). Natural "Natural Experiments" in Economics. Journal of Economic Literature, 38(4), 827-874.

Slonim, R., Wang, C., Garbarino, E., & Merret, D. (2012). Participation Biases in the Lab. Working Paper.

Smith, V.L. (1962). An experimental study of competitive market behavior. Journal of Political Economy, 70(2), 111-137.