The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

On the Generalizability of Experimental Results in Economics: With a Response to Commentors

Omar Al-Ubaydli
John A. List

CESIFO WORKING PAPER NO. 4543
CATEGORY 13: BEHAVIOURAL ECONOMICS
DECEMBER 2013

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4543*

On the Generalizability of Experimental Results in Economics: With a Response to Commentors

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-Code: C900, C910, C930.

Keywords: lab and field experiments, generalizability.

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

John A. List
Department of Economics
University of Chicago
USA - 60637 Chicago IL
jlist@uchicago.edu

November 2013
We wish to thank Colin Camerer, Marco Castillo, Robert Chambers, David Eil and Andreas Ortmann for helpful comments and discussions. Alec Brandon and David Novgorodsky provided excellent research assistance. This study extends NBER w17957 “On the generalizability of experimental results in economics,” by including a discussion of studies that have commented on our past work.

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.

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 eloquently quipped nearly 100 years ago: 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, Camerer, Fréchette, Kessler and Vesterlund this volume). 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 (2007b; 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, this volume represents 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

[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.

behaviors aggregate to importantly affect market equilibria, and how does equilibration affect the individual behaviors?

The object of this forum is to discuss the recent study due to LL. For the most part, the critics writing in this volume understood LL's contributions and hypotheses, and wrote a balanced and thoughtful evaluation of that work. As a point of reference, 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, as pointed out by some of the commentators in this Volume). We suspect that the commentators are in full agreement with this broader point.

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 then place the theory into a convenient 'updating' model that shows the power of replication: we argue that just a few replications yields a tremendous increase in the probability that the received result is in fact true. Our penultimate section addresses some of the various criticisms leveled by the discussants. 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.

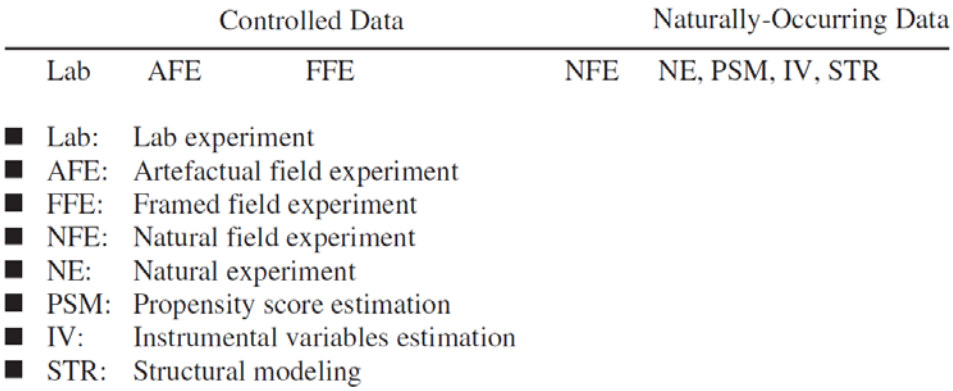

Figure 1: A field experiment bridge

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 quite 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 explored in Fréchette (this volume). As Fréchette discusses, 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 (2011). 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 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 volume. 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:

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

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

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 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. 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.

[4] Various people use the term external validity. As we noted in Harrison and List (2004, p. 1033), 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."

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 $\bar{h}$:

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

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 $\mathcal{F}^0 = \left\{F^0_{x,x',z} : (x, x', z) \in S_X \times S_X \times S_Z\right\}$ 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:

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

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.

[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.

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 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 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 to 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 (including Camerer in this volume) 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.

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

[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.

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

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 are 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 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. 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 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

[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.

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 (x, x', z) where we vary two of the dimensions of z. The space is divided into natural environments (above the dashed line) and non-natural environments (below the dashed line). One combination of (z_1, z_2) 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.

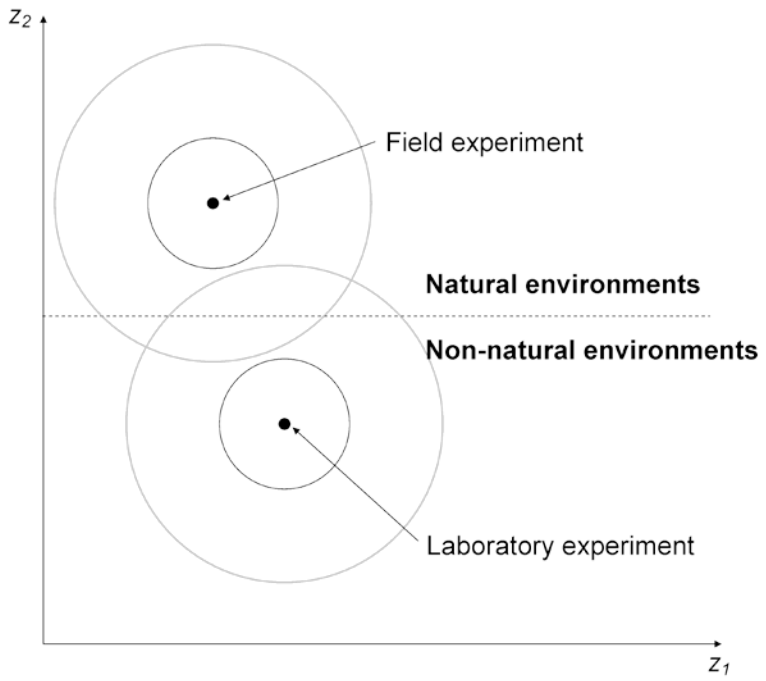

Figure 2: Generalizability in field and lab experiments

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 T 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 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:

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

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,

[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.

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. 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

[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.

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. (2007). 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 is another important virtue of laboratory experimentation—the ease of replication. Since replication is the cornerstone of the experimental method, it is important to discuss briefly the power of replication. Although such tracks have been covered recently by Maniadis, Tufano and List (MTL 2013), we parrot their discussion here because there has been a scant discussion of this important issue and it serves to highlight a key virtue of laboratory experimentation.

As Levitt and List (2009) discuss, 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. Camerer does this with List's (2006a) data, as have dozens of other scholars. Similar to all of the other scholars, when analyzing the data Camerer finds results that are identical to what List (2006a) reported.

A second notion of replication is to run an experiment which 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.[12] We show how replication and generalizability are linked using the theoretical model and our empirical example is on the second notion of replication, or whether use of the exact same protocol leads to similar empirical results. Yet, our fundamental points apply equally to the third replication concept.

[12] This taxonomy is in agreement with the one proposed by Cartwright (1991). Hunter (2001) also defines three levels of replication, but the first level he suggests concerns the exact repetition of the original study in all dimensions, rather than the routine checking of the original study. His other two levels are largely equivalent with ours.

Continuing with the notion that the researcher has a prior on the actual behavioral relationships, we follow MTL's model of replication. 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.[13] Let α denote the typical significance level in the field (usually $\alpha = 0.05$) and $1 - \beta$ denote the typical power of the experimental design.[14] We are interested in the Post-Study Probability (PSP) that the research finding is true. Or, more concretely, given the empirical evidence, how sure are we that the research finding is indeed true.

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

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

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

For our purposes, we can use [1] to determine the reliability of an experimental result. We plot the PSPs under three levels of power in Figure 3 (from MTL).

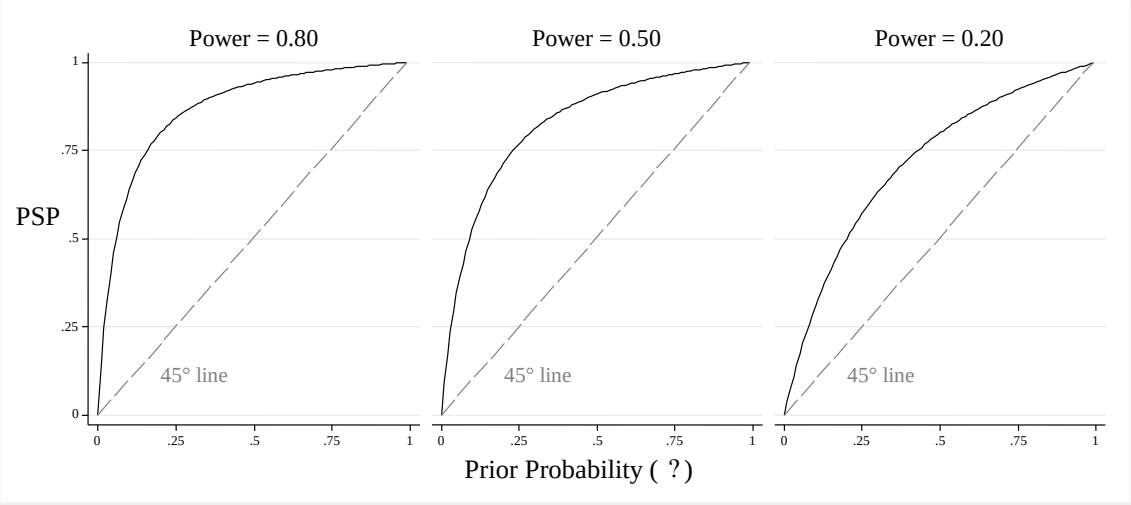

Figure 3: The PSP as a function of power

[13] 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, $\pi = \Pr\{H_1 \; is \; true\}$.

[14] As List et al. (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?

Upon comparing the leftmost and rightmost panels for the case of $\pi = 0.5$, we find that the PSP in the high power (0.80) case is nearly 20 percent higher than the PSP in the low power (0.20) case. This suggests that as consumers of experimental research, we should be nearly 20% more certain that research findings from higher powered experiments are indeed true in comparison to lower powered experiments. What else Figure 3 tells us is that we should be wary of 'surprise' findings of (those that arise when π values are low) from experiments because they are likely not correct findings if one consider the low PSP. In this case, they are not even true in the domain of study, much less in an environment that the researcher wishes to generalize upon.

Figure 3 powerfully illustrates the importance of replication for our scientific endeavors. Society can have much more confidence in our results as long as a few replications take place. The insights in Figure 3 are inspired by a recent study due to Moonesinghe et al. (2007), who show that even a handful of replication studies may suffice to draw the correct inference about the true association.

Of course, there are several other important virtues of laboratory experimentation, and Levitt and List (2007a, 2007b) point them out. Most importantly, LL (2007b) note: except in rare circumstances, laboratory experimentation is likely to be a useful tool for providing *qualitative* evidence, even when the generalizability of deep structural parameters is suspect.

4. Reflections on commentary

The Editors of this volume have asked for, and received, very important discussions of the generalizability issue. We admire the work of all four authors. Fréchette (2012) importantly explores whether students and experts behave similarly in lab experiments. Implicitly the paper contributes to our understanding of lab treatment distributions across populations. His paper represents a great service piece, as it takes on the invaluable question: would one reach similar conclusions using professionals as opposed to the standard subject pool of undergraduate students?

When considering a comparison of behavior across subject pools, three important features arise: preferences, selection rules into the experiment, and behavioral effects of taking part in an experiment. To illustrate why the three features are important to consider, let's first assume that the true treatment effects $\tau_i = y_{i1} - y_{i0}$ are identical across subject pools. That is, one would obtain identical treatment effects using a NFE.

What should we find when we compare a laboratory experiment with an AFE? We very well might find significant population differences. Even in the case that the treatment effects in the student and expert population exactly overlap, selection into the experiment and scrutiny effects might cause them to look different.

Likewise, even in cases when the observed data across students and experts are statistically indistinguishable, selection or the effects of scrutiny might be masking true differences. These are potentially important aspects to consider when comparing results across groups of people, as pointed out in LL:

> One approach to investigating subject pool biases is to examine whether professionals, or other representative agents, and students behave similarly in laboratory experiments. Fehr and List (2004) examine experimentally how chief executive officers (CEOs) in Costa Rica behave in trust games and compare their behavior with that of Costa Rican students. They find that CEOs are considerably more trusting and exhibit more trustworthiness than students. These differences in behavior may mean that CEOs are more trusting in everyday life, or it may be that CEOs are more sensitive to the lab and non-anonymity effects discussed above, or that the stakes are so low for the CEOs that the sacrifice to wealth of making the moral choice is infinitesimal (Levitt and List 2007b: p165-166).

What this means is that without the help of a selection and behavioral model, it is difficult to interpret data from simple comparisons of experts and students. Creating models that predict when behavior will be different, why it is different, and what factors exacerbate or attenuate these differences helps us explore questions such as "Is behavior of students (or experts) who select into the lab representative of behavior of experts in the field?" and "Does a given treatment affect people in the experiment the same way it affects people in the population at large?" In the end, we are after how people behave in natural settings. Yet, experimenting in important parameter spaces in the laboratory can help us tremendously in understanding behavior in natural settings. Fréchette (2012) takes us in this direction in a very welcome manner.

Kessler and Vesterlund (2011) is also a clear and thoughtful discussion of experimentation. Importantly, they concisely point out the most important misunderstood argument in LL when they note (p. 1):

> While the debate has centered on the extent to which the *quantitative* results are externally valid, we will argue that for most laboratory studies it is only relevant to ask whether the *qualitative* results are externally valid. Interestingly, among the authors on both sides of the debate there is significantly less (and possibly no) disagreement on the extent to which the qualitative results of a laboratory study are externally valid.

One of the main ideas in Levitt and List (2007a) was that the lab did not provide a good means to measure deep structural parameters. That is, LL questioned the integrity of the point estimates—for example, 63% of people have strong altruism—from laboratory exercises. The original paper discussed the various games and concluded that proper inference of results from games, such as the dictator game, had been mistaken. Studies by and large have shown the fragility of such results. During an early debate on the LL study at the ASSA meetings in Boston in January 2006, interestingly, Camerer and others on the panel strongly refuted this point. In fact, their beliefs had recently been summarized in published work arguing that the dictator game is a useful tool to *measure* social preferences (Camerer and Fehr 2004):

> Regardless of which models are most accurate, psychologically plausible, and technically useful, the important point for social scientists is that a menu of games can be used to measure social preferences, like the extent to which people weigh their monetary self-interest with the desire to reciprocate (or limit inequality), both negatively (in ultimatum games) and positively (in trust games), and with pure altruism (in dictator games).

We are happy to learn that Camerer (2011) has seemingly changed his mind and now notes that:

> It is true that early discussions called the dictator game a way to measure "altruism," as compared to generous ultimatum offers by selfish proposers which are meant to strategically avoid rejection (e.g. Eckel and Grossman, 1996; Camerer, 2003). LL (2007a, Table 1) repeated this standard view, describing the "social preference interpretation" of dictator game giving as "altruism; fairness preferences, such as inequity aversion." The idea that dictator giving results from impure altruism, either warm glow or a preference for a social image of adhering to sharing norms, arose in the mid-1990s. The social image account was noted early on by Camerer and Thaler (1995), who used the phrase "manners." Evidence accumulated in the last few years is also consistent with aspects of warm glow and social image motive (i.e., appearing to have good "manners").

Taken together, this is a good sign that researchers are accounting for accumulated empirical evidence and changing their views accordingly. By and large, we agree with the major points of Fréchette (2012) and Kessler and Vesterlund (2011). In numerous cases, however, we disagree with Camerer's (2011) reading of our work and his interpretation of experimental work more broadly. In the interests of parsimony, we focus on the most important points of disagreement.

A first point that Camerer makes is that there is nothing inherent with features of the lab that limits their generalizability (p. 4):

> We then consider which common features of lab experiments might threaten generalizability. Are those features a necessary part of all lab experiments? Except for obtrusive observation in the lab—which is an inherent result of Federally-regulated human subjects protection in the US (though not in all countries)—the answer is "No". The special features of lab experiments which might limit generalizability can therefore be relaxed, if necessary, to more closely match particular field settings. Then we ask whether typical lab features necessarily undermine generalizability to all field settings in a way that cannot be corrected. They do not.

As shown above within two different theoretical frameworks, it is straightforward to see how this statement is incorrect. Indeed, we have noted several times in the literature that NFEs deliver different treatment effect estimates than the lab, AFEs, and FFEs deliver. In a recent study, in fact, List (2011) points this out quite directly when he notes:

> One possible source of bias in these other experimental approaches is that generally the subjects who choose to participate in the experiment are those who expect to gain the most (perhaps because they believe they are likely to get good results from the treatment). As a result, the estimated causal effect from these other experimental types, while valid, might not generalize to the target population of interest—which of course includes the subpopulation (often the majority) that did not volunteer for the experiment when offered the opportunity.
>
> Natural field experiments address this problem. Because subjects do not make a choice about whether they will participate, the treatment effect obtained from natural field experiments is, in the best-case scenario, an estimate that is both causal and broadly generalizable (in Al-Ubaydli and List, 2012, my coauthor and I offer a formal treatment).

> Put simply, since participants in the natural field experiment are a representative, randomly chosen, non-self-selected subset of the treatment population of interest, the causal effect obtained from this type of experiment is the average causal effect for the full population—not for a nonrandom subset that choose to participate (List 2011: p6-7).

Accordingly, in direct contrast to Camerer's argument, there *are* typical features of lab (and AFEs and FFEs) experiments that threaten generalizability. We hope that this study and our previous work can finally resolve that misunderstood issue.

One should then ask if there is important empirical evidence that shows the selection effect outlined above is important. LL present some evidence on both sides of the fence, and we point the reader to LL for further discussion of those studies. In our own field work, we have not found considerable evidence of selection being important in our FFEs. But, a recent innovative paper—Slonim et al. (2012)—does report strong evidence of selection being an important issue within their laboratory experiments.

Slonim et al. (2012) investigate the representativeness of participants in economics lab experiments as part of the greater research agenda of testing and interpreting the generalizability of lab experiments in economics. In particular, while a number of studies have touched upon the variance in pro-social and other behaviors within experiments by using relevant subject-level observables, there has been less emphasis on comparing the individuals that were invited to the experiment to those who eventually participated.

Slonim et al. study participation in economics experiments from a standard labor supply perspective, and using an expected utility framework, derive a set of hypotheses across a number of different subject-level dimensions. Four of the hypotheses generalize to all lab participants and are as follows (all stated relative to non-participants): lab participants are predicted to have less income, more leisure time, greater interest in lab experiments and greater academic curiosity (particularly those with revealed interest in economics topics), and greater pro-sociality in terms of volunteering time.[15] The remaining hypotheses are dependent on the parameters of the experiment itself. Specifically, the greater the uncertainty in the payoffs, participation conditional on agreeing to participate, length of the experiment, etc., the more likely that lab participants are to be (relatively) less risk-averse. In contrast, having subjects participate by appointment only will increase the participation of risk-averse individuals relative to a "drop-in" economics experiment.

To test these hypotheses, Slonim et al. begin with a sample of 892 students in an introductory economics class which is split across 77 one-hour tutorial sections taught by 22 tutors. In the fourth week of classes, students were asked to voluntarily complete three incentivized experiments and a 20 question survey, and 96% obliged in their tutorial and after completing these, the students were presented with a random flyer advertising a lab experiment that the students could participate in anywhere between 7 – 13 days after the above tutorial section. Slonim et. al also randomize the participation requirement over the following "appointment type" treatments: 1) by appointment only, 2) by "drop-in," or 3) by either, i.e., the subject's choice.

[15] The authors suggest that this is distinct from pro-social behavior in terms of monetary decisions which preliminary findings suggest is correlated with wealth, introducing countervailing forces into the participation decision.

Looking at the main predictions of Slonim et al., there is strong support for all four of their main hypotheses in that lab participation is decreasing in spending per week (a proxy for student income; controlling for household income and hours worked), decreasing in work hours per week (used to estimate the remaining hours that could be used for leisure), increasing in interest in economics[16] and ability to make consistent decisions (both proxies for academic curiosity), and increasing in the number of times an individual volunteers a year (though not in total volunteering hours).

Additionally, they find support for the hypotheses that more risk-averse individuals choose to make an appointment when given the choice and that the propensity to save (as a proxy for patience) is predictive of participation in the lab. Using the full model and starting with a simplifying assumption that all of the above qualities are equally distributed in the population, they find overrepresentation of over 2 to 1 for the relevant income, leisure time, intellectual interest, and pro-sociality measures.

Slonim et al. also touch upon a number of suggestions for optimal design to alleviate some of the issues of non-representative samples. For example, to increase the representation of higher income subjects, experimenters could increase the experimental payments or reduce the length of the experiment (for those whose high income comes not through household wealth but hourly income). Similarly, excluding mention of economics or various social value frames can help solve the issue of overrepresentation of these types in the experimental subject pool. Alternatively, experimenters could collect the relevant observables irrespective of participation and then control for these in reporting their results. For example, when adding individuals to a mailing list to distribute information about upcoming experiments, experimenters could ask potential subjects to complete a full survey measuring the qualities that may otherwise bias results. Yet, one should still recognize that unobservables might differ across participants and non-participants.

The overarching point of Slonim et al.'s important contribution is that treatment effects in lab experiments can be deceiving because generalizing from experimental findings can misrepresent bias in the treatment effects. When explicitly stating an average treatment effect within an experiment, this bias only presents a problem to the extent that the particular non-representativeness impacts the relevant outcome measure independently of the treatment effect. However, when looking to generalize the experimental results to economic agents in naturally occurring markets, which include participants and non-participants, a host of issues arises if the participation decision correlates with observable qualities of individuals that correlate with the outcome measure independently of the true underlying treatment effect.

In his own description of Slonim et al. (2012), Camerer himself concedes this point, and he suggests three remedies: using selection models, increasing the participation rate by increasing the show-up fees, and studying groups that ex ante have very high participation rates. We would regard these remedies as ineffective in a wide range of environments studied by economists, especially when we take into account the other advantages that we associate with natural field experiments.

Sorting of course is not merely a lab phenomenon. Della Vigna et al. (2012) highlight the importance of sorting in the field. Indeed, very rarely do economic outcomes in the field not have a participation decision associated and recent field research highlights the importance of taking sorting into account. For

[16] The authors note that given that the experiment is advertised as an "economics decision-making" task, it is difficult to rule out this potentially differential marketing as increasing this group's participation.

example, Gautier and van der Klaauw (2010) examine pay-what-you-want choices by visitors to a hotel that either knew or didn't know that the hotel was pay-what-you-want before making a reservation. Visitors that knew that the hotel was pay-what-you-want ex ante paid significantly less than those that did not, suggesting that a different sample sorted into visiting the hotel when the information was posted.[17]

Empirical Evidence

Our final point concerns Camerer's discussion of the studies that most closely match the lab and the field. Here, beauty is certainly in the eye of the beholder because our interpretation of our own data and others' data certainly departs from Camerer's interpretation.

Whether you like the List (2006a) study or not, as the reader learned from Camerer's summary of our work, when we have built a bridge between the lab and field, our own data (see, e.g., List 2004, 2006a, 2009, and the papers cited below) sometimes find good generalizability, and in others finds that the lab does not provide a good indication of behavior in the field—especially when measuring quantitative insights. We detail a few of those studies now and discuss some of the work that Camerer cites that is closely related to our own work.

We begin with the re-analysis of the data in List (2006a). Camerer is one of many scholars who have asked for the List (2006a) data (another prominent scholar who has asked for (and thoroughly analyzed) the data is Ernst Fehr; many graduate students have also scoured the data). Each one of them has delivered replication in the first sense of the definition above: the data show what List (2006a) reported. In addition, all of them but Camerer have been satisfied with the conclusions drawn (or at least we have not heard of any dismay). To avoid a tiresome rehashing of List (2006a), we will provide a very focused response to Camerer which necessarily assumes some familiarity with List (2006a) on the reader's part. Those interested in a fuller treatment need only read the original paper (List 2006a) since we are adding nothing of substance.

A first fact (as Camerer notes) is that quality of the good delivered is much lower in the field than in the lab. Even though the same prices are paid, when dealers know that they are taking part in an experiment, they deliver higher quality. This is Camerer's first finding (p. 29), and we agree, that is what the data suggest. This is a strong pattern that is difficult to get rid of in the data, even when you try. One could stop here and the lab-field generalizability point is answered—in this setting, when a person knows that they are in an experiment they provide higher quality. But, the data are much richer.

The crux of the other findings in the paper starts with the fact that in the field, local dealers have weakly stronger reputational concerns than non-local dealers do. In the field, in order of weakly increasing reputational concerns, we have periods when (1) there is no grading and no announcement of the intent to grade, (2) there is no grading but there has been an announcement of the impending arrival of grading, and (3) there is grading. There are therefore six configurations of field reputational concerns: local vs. non-local and no-grading vs. announced grading vs. grading. The key is that one can weakly rank in each dimension.

[17] Researchers have developed techniques to deal with sorting, see the innovative study of (Lazear et al. 2012).

In the laboratory, local and non-local dealers exhibit anonymous gift giving. In multiple reputational configurations in the field, zero gift giving is observed, supporting the claim that there are zero (or weak) social preferences. Comparing behavior across reputational configurations in the field, when reputational concerns are weakly higher, anonymous gift giving is weakly higher, and in certain configurations, gift giving is very strong. This bolsters the argument that the laboratory behavior of dealers is not caused by social preferences and that it is indeed impossible to eliminate reputational concerns in the laboratory.[18]

Unsatisfied, Camerer asked for the data so that he could conduct his own investigation. He starts by pointing out that in certain configurations, gift giving in the field is stronger than anonymous gift giving in the laboratory. Upon reading footnote 18 and List (2006a), and understand the design and the point that is being made, it is unclear what the value of Camerer's finding is—neither List (2006a) nor us would refute this new 'finding,' but we find it troublesome that it is even a claim. It really has no significance whatsoever given the dimensionality of the design.

Further, there is nothing to suggest that List's conclusions are driven by a lack of power in the statistical test—one just needs to look at the figures in the original study to learn about the behavioral patterns. Despite this, Camerer curiously argues that in the quest for power, he wants to discard a large portion of the data and to focus on a small subset of the data. Camerer also claims that "it is the closest match of lab and field features," again misunderstanding the point of List (2006a).

Camerer conducts some additional tests on these data and finds that if you use some relatively unconventional tests (rather than those that List uses), some of the results switch from significance to marginal significance, and that therefore, all bets are off. (As an aside, according to our attempts at replicating what Camerer did, we obtain different results for two of the three tests, though the thrust of his argument is not affected substantively.)

We describe two of the tests (Epps-Singleton and Fligner-Policello) as relatively unconventional since a Google Scholar search of papers written by Camerer where the names of these tests appear in the text yields four papers in the case of E-S and one paper in the case of F-P out of dozens and dozens of total papers published (Colin is a prolific writer). This is particularly puzzling since his supposed justification is the non-normality (in the Gaussian sense) of List's data; surely non-normality is a regular occurrence in Camerer's own data since much of it is experimental and of the same nature as our own.

Indeed, a Google Scholar search of the *American Economic Review* finds one use of the Epps-Singleton test and zero uses of the Fligner-Policello test, and analogous results are obtained for the *Quarterly*

[18] Camerer also mentions the Levitt and List (2008) *Science* paper. The reader might also wish to read that study, as it is a 1-page summary of the use of behavioral economics and field experiments. The point of the figure was to show that when the conditions are right, observability will affect behavior. We could have easily shown Camerer's first fact, discussed above. Or, we could have easily shown the mendacious claims data, the local versus non-local dealer sportscard data over the third-party verification period, or the ticket stub data from both the local and non-local dealers. But, as evidence, we showed the data for the non-local dealers across the lab and field. They all showed the same behavioral pattern. Camerer thinks that we should have shown the local dealer data across the lab and field. This is quite puzzling. If he would have read the List (2006a) paper carefully, he would have known that was the point of the entire exercise: local dealers should show signs of gift exchange in both the lab and field, thus there should not be differences in behavior across domains because reputational concerns are driving them to engage in gift exchange (strategic reciprocity is at work)!

Journal of Economics, *Econometrica*, and the *Journal of Political Economy* (the first and only four journals that we checked).

Given the infrequency with which both Camerer and the literature use these tests compared to the usual array of statistical tests that we and the literature use, we regarded conducting rigorous Monte Carlo simulations to compare power as unproductive. Instead, we conjecture that most statistically significant results in economics will become marginally significant if you drop enough of the data and attempt a large enough number of statistical tests.

Camerer also chose not to comment at all about the data that he chose to throw out of the analysis. Does he think that they are useless data? For example, how does Camerer explain the finding that prior to any announcement of grading, *both* local and non-local dealers exhibited zero gift giving (columns 3 and 4 in Table 5 of List 2006a)? This is hardly an afterthought—it is very much the crux of List (2006a). In fact, a key to the design of List (2006a) was that it contained several outcome measures across several settings to measure behavioral change and link it to theory. Discarding hundreds of observations and focusing on a handful (as Camerer does) seems too demanding of the original data given the richness of the original design.

There is mounting evidence that the main results on generalizability found in List (2006a) are replicable and quite prominent in every setting. Winking and Mizer (2013) design laboratory and field versions of a dictator game. The goal is isolating scrutiny/knowledge of being in the experiment as the only difference between the two versions, in a similar fashion to List (2006a). The subtlety of their experimental design means that it is too long to warrant being reproduced here. However, from the perspective of Camerer's comments on List (2006a), it is reassuring to find virtually identical results: in the laboratory version of the dictator game, behavior consistent with the plethora of existing laboratory studies was observed, i.e., substantial donations. In contrast in the field version, every single participant opted to donate nothing, choosing instead to keep all the money for themselves. Likewise, Yoeli et al. (2013) use a natural field experiment to show the observability effect (p. 1): "We show that observability triples participation in this public goods game. The effect is over four times larger than offering a $25 monetary incentive, the company's previous policy." They proceed to note that "People are substantially more cooperative when their decisions are observable and when others can respond accordingly," and cite 20 recently published studies as further evidence.

Ultimately, the content of this exchange on List (2006a) is irrelevant to our big picture arguments about the pros and cons of field experiments. But, to complete the discussion, we now turn to a brief discussion of a few of the other studies discussed by Camerer. We are not interested in a food-fight here, so we briefly comment on these strands of work. The interested reader should certainly read these other studies, the Yoeli et al. (2013) work, and the citations therein, and come to an informed opinion on his/her own.

Donations, fishermen, and soccer

A related study (Benz and Meier 2008) that Camerer discusses was actually published in a special issue volume that List edited on field experiments for the journal *Experimental Economics*. The Benz and Meier (2008) study permits a novel test of generalizability of results across domains, and presents results that are quite consonant with List (2006a). Here is what List (2008) wrote about the study in his introduction, and as far as we know the data have not changed since the paper was published:

> The study takes advantage of a naturally-occurring situation at the University of Zurich, where students are asked to give money towards two social funds. The authors undertook a framed field experiment by setting up donation experiments in the lab that present students with the identical field task of giving to two social funds, but wherein they know that they are taking part in an experiment (and their endowment is provided by the experimenter).
>
> The authors are therefore able to construct a unique data set by using panel information on charitable giving by individuals both in a laboratory setting and in the field. In this respect, different from Eckel and Grossman (2008) and Rondeau and List (2008), Benz and Meier are able to change the properties of the situation without changing the population.
>
> In making comparisons across these decision environments, Benz and Meier (2008) find important evidence of positive correlation across situations, but ultimately find that giving in the lab experiment should be considered an upper bound estimate of giving in the field: subjects who have never contributed in the past to the charities gave 75 percent of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment.

Importantly, *subjects who have never contributed in the past to the charities gave 75 percent of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment*. In short, these data paint a very similar picture to what List observed amongst his sportscard dealers—the nature and extent of scrutiny affects behavior in predictable ways. This result is also found in a recent study due to Alpizar et al. (2008), who find that scrutiny is an important determinant of pro-social behavior. When done in public, charitable gifts increase by 25%, suggesting the power of scrutiny. List (2006b) provides a discussion of many more examples in this spirit.

Camerer also discusses a recent study of fishermen, due to Stoop et al. (2012). The authors present a clever assortment of experimental treatments to explore cooperation rates and why they arise in the manner in which the literature has discussed. They begin by conducting a FFE measuring cooperation among groups of recreational fishermen. They carefully construct the setting to mimic the classic Voluntary Contributions Mechanism (VCM). Interestingly, they find little evidence of cooperation, even though the received VCM lab results represent one of the most robust performers in delivering favorable cooperation rates.

They do not stop there, however. In an effort to learn why such departures occur, they build an empirical bridge in the spirit of List (2006a; 2009) to identify the causes of the behavioral differences. They rule out the subject pool and the laboratory setting as potential causes of behavioral differences. The important variable within their environment that causes behavioral differences is the nature of the task. Importance of the nature of the task is consonant with Gneezy and List (2006) and Harrison et al. (2007). As Harrison et al. (2007) put it: such results highlight that the controls that are typically employed in laboratory settings, such as the use of abstract tasks, could lead subjects to employ behavioral rules that differ from the ones they employ in the field. Because it is field behavior that we are interested in understanding,

those controls might be a confound in themselves if they result in differences in behavior (we discuss Harrison et al. 2007 below).

This paper is important in that it provides a reinforcement of a broader point lying beneath other studies exploring behavior in quite different domains: it shows that games that have the same normal form can generate very different behavior, and illustrates how important this can be when making laboratory and field comparisons. While Stoop et al. (2012) are able to obtain basically the same result (with modest differences) in the lab and field when the game was conducted in the same way, they find a dramatic difference in behavior when they change the nature of the task of the game, even if they change it in a way that is inconsequential in theory. This means that, for this specific case, they can isolate the exact effect leading to the lab/field difference.

Finally, Camerer also argues that the lab has advanced our understanding of game theory by testing whether subjects can play minimax in zero-sum games with a unique mixed strategy equilibrium. Minimax yields a set of stark predictions for optimal play in such games, but decades of lab experiments (see Levitt et al. 2010 for a discussion) have found that subjects consistently deviate from optimal play, even when directly instructed to do so (Budescu and Rapoport 1992).

Palacios-Huerta and Volij (2008) test minimax in a 2x2 and 4x4 game and find that the typical lab subjects do not play minimax, but in an AFE subjects with experience (in particular kicking penalty shots in soccer) play just as minimax would predict. What is awkward about Camerer's choice of citing Palacios-Huerta and Volij (2008) is that this particular study finds that the standard lab experiment does not predict well when it comes to understanding behavior in AFEs or in the field. In fact, the only sample that Palacios-Huerta and Volij (2008) argue fails game theoretic predictions are undergraduate students—that is, the only sample they find that does not play minimax is students in their lab experiment!

Nonetheless, this result showed be viewed with great caution. The contribution of Palacios-Huerta and Volij (2008) to our understanding of lab or field behavior is unclear, as subsequent attempts at replication have failed (see MTL, discussed above).

In particular, Levitt et al. (2010) test whether professional soccer players, poker pros, and bridge pros can transfer their experience in the field into a lab-type setting and none do. What Levitt et al. (2010) find is that whether it is students or professionals, when they play lab games *none of them behave in accord with minimax predictions*. Thus, *if* professionals do play minimax in the field, they do not transfer that behavior to the lab games that Levitt et al. (2010) employed.

Whether actual behavior follows the predictions of minimax is still unclear and what seems certain is that more lab and field experiments are necessary. Perhaps with big enough stakes[19] and experienced enough pros evidence could be produced confirming minimax in a simple 2x2 game.

[19] Note that this particular effect remains an open question in the economics literature. Both Smith and Walker (1993) and Camerer and Hogarth (1999) find support for the use of financial incentives over hypothetical stakes, but both suffer from opportunistic samples of existing literature. One notable exception can be found in Hertwig and Ortmann (2001) where, using a 10-year sample of studies published in the *Journal of Behavioral Decision Making*, the authors find that financial payments may improve task performance and decrease outcome variance, but also call upon using financial payments as explicitly independent variables in future experimental studies to provide a direct empirical test.

Open air markets

List (2009) uses open air markets as a natural laboratory to provide insights into the underlying operation of such markets (this discussion closely follows List 2009). His approach reformulates and extends the problem of stability of equilibria, moving it from an abstract theoretical concept to a question about the behavior of agents in an actual marketplace. A first key result from the field experiments is the strong tendency for exchange prices to approach the prediction of competitive market equilibrium theory. Even under the most severe tests of neoclassical theory (treatments that predict highly asymmetric rents) the expected price and quantity levels are approximated in many market periods. Consonant with the above discussion on market experiments, these results suggest that in mature markets very few of the "typical" assumptions, such as Walrasian tâtonnement or centrally occurring open outcry of bids and offers, are necessary to approximate the predicted equilibrium in the field (see also List 2004).

Yet, such markets are ripe for price manipulation. For instance, in certain cases small numbers of sellers provide homogeneous goods that are jointly purchased from middlemen, certain barriers to entry exist, and seller communication is continual. Indeed, during the course of conducting the original tests of neoclassical theory List learned interesting details of just such conspiracies in these markets. Armed with knowledge of a mole (confederate) he built a bridge between the lab and the field, effectively exploring the behavior of experimental subjects across sterile and rich settings. This approach has a dual benefit in that it affords an opportunity to marry the vast experimental literature on collusion in laboratory experiments (see, e.g., Holt 1995 for an excellent overview) with parallel behavior in the field.

Accordingly, he began with a general line of attack to undertake controlled experiments where factors at the heart of his conjectures were identifiable and arise endogenously. He built a bridge by beginning with data generation from a controlled laboratory study with student subjects. He proceeded to collect data using the exact same protocol with subjects from the open air market (denoted "market" subjects below) who have selected into various roles in the marketplace. He then executed a series of controlled treatments that slowly moved the environment from a tightly controlled laboratory study to a natural field experiment. By slowly moving from the typical laboratory setting to the naturally-occurring environment, he is able to ascertain whether behavior differs across the lab and field domains. And, if so, he can determine the important features that drive such departures.

In this regard, comparisons of interest include observing behavior of (i) identical individuals in the lab and the field, (ii) agents drawn from the same population engaged in lab and field experiments, where the lab selection rules might be different from the way in which markets select individuals, and (iii) individuals drawn from different populations engaged in lab and field experiments.

List reports that individuals drawn from different populations show considerable signs of an ability to maintain collusive ties, even in austere situations. Yet, there are some behavioral disparities. For instance, students are influenced much more by changes in anonymity whereas marketers are influenced to a greater extent by context. For the case of agents drawn from the marketing population and placed in lab and field experimental roles, he reports only marginal evidence that selection is important. For example, when comparing cheating rates in the natural field experiment across those who agreed to participate in the lab experiments and those who refused, he finds little evidence of significant differences. Though we should highlight that this comparison is made with a sample size of 17 sellers

who agreed to participate in a controlled lab or framed field treatment, and 5 sellers who turned down the request (5 sellers were never asked) but participated (unknowingly) in the natural field treatment.

Finally, examining the behavior of the 17 individual sellers who were in experiments across the lab and the field provides insights into generalizability of results across domains. Levitt and List (2007b) argue that being part of an experiment in and of itself *might* induce certain types of behaviors. In the current case, one hypothesis is that conditional on making collusive arrangements, taking part in an experiment *might* induce sellers to more readily maintain their collusive promises. More broadly, the conditions set forth in an experimental situation might induce behavioral tendencies that are simply not observed in the field.

Using data from the 17 sellers (11 of whom were in a lab treatment, 3 of whom were in a lab and framed field treatment, and 3 of whom were in a framed field treatment), he reports a small correlation between cheating rates across the lab treatments with no context and the other environments. However, in a simple regression model, the best predictor of whether sellers will cheat in the natural field experiment is their measured cheating rate in the framed field treatments and the lab treatment with context.

Figure 4 highlights one aspect of List's results—rates of cheating across the experimental domains. Whereas cooperation rates are found to be quite high in the tightly controlled framed field experiments, he finds that when sellers do not know that they are part of an experiment they cheat on the explicit agreements much more often. The level of cheating observed in the natural field experiment is larger than cheating rates observed in any of the tightly controlled framed field treatments. However, in *aggregate* the Figure shows that the best predictor of cheating rates in the natural field experiment is behavior in *sterile* laboratory treatments with neutral language instructions.

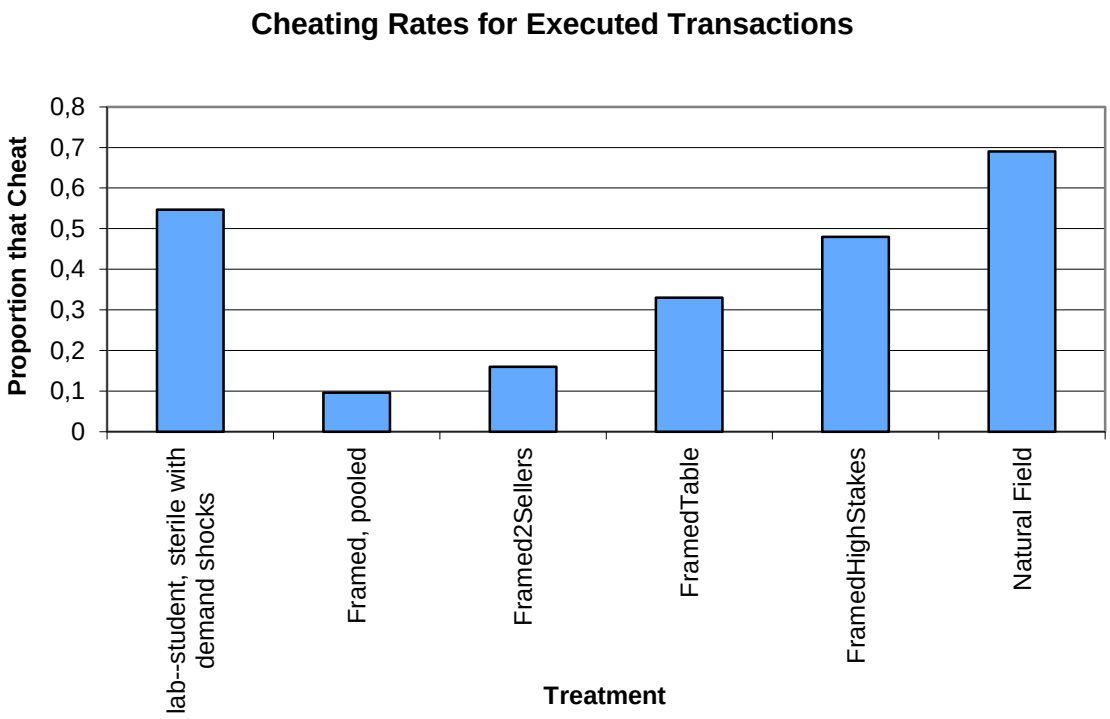

Figure 4: Cheating rates for executed transactions across the various treatments
Figures provide the proportion of actual transactions that were at prices below the agreed upon collusive price.

Other Work

Rather than walk through every study that is published in this area we can conclude that some lab work shows good signs of generalizability, and others do not. We should continue to explore empirically theoretical structures such as the framework presented above to learn more about this first order question. In this spirit, our model highlights that one area that has been largely ignored in this discussion is the fact that the typical context in which laboratory experiments are completed impose artificial restrictions on choice sets and time horizons.

One illustration of this is the work of Gneezy and List (2006), who observe work effort in two jobs where some employees are randomized into a gift treatment. The typical lab session does not exceed an hour, yet the typical labor supply decision is not made over the course of an hour. Gneezy and List find that over the course of six hours of work the gift increases worker productivity, but by the fourth hour the impact of the gift on behavior was nil. Camerer (2011) argues that the similarities of gift exchange in the lab and field during the first hour are actually evidence of the generalizability of results from the field.

Yet this more modest version of gift exchange is hardly what Fehr et al. (1993) had in mind when, in their abstract, they concluded that "These results provide, therefore, experimental support for the fair wage-effort theory of involuntary unemployment". No doubt, it is possible to do a multi-hour, real effort experiment in the laboratory—one that would cleanly examine the robustness of Fehr et al.'s results, but nobody did, confirming the original interpretation of Fehr et al. (1993).

Another illustration of how important this issue can be is found in Harrison et al. (2007). Their abstract tells the complete story:

> Does individual behavior in a laboratory setting provide a reliable indicator of behavior in a naturally occurring setting? We consider this general methodological question in the context of eliciting risk attitudes. The controls that are typically employed in laboratory settings, such as the use of abstract lotteries, could lead subjects to employ behavioral rules that differ from the ones they employ in the field. Because it is field behavior that we are interested in understanding, those controls might be a confound in themselves if they result in differences in behavior. We find that the use of artificial monetary prizes provides a reliable measure of risk attitudes when the natural counter- part outcome has minimal uncertainty, but that it can provide an unreliable measure when the natural counterpart outcome has background risk. Behavior tended to be moderately risk averse when artificial monetary prizes were used or when there was minimal uncertainty in the natural nonmonetary outcome, but subjects drawn from the same population were much more risk averse when their attitudes were elicited using the natural nonmonetary outcome that had some background risk. These results are consistent with conventional expected utility theory for the effects of background risk on attitudes to risk.

More importantly, the point that studies such as these make is that abstracting the field into the lab is not easy, and key elements of decisions made in the field, like length of work or the behavioral features that might guide decisions in the field, are missing from typical laboratory experiments.

Another example of this is Dahl and DellaVigna (2009) which compares lab and field evidence on aggression and violent movies. What previous lab experiments failed to consider in their studies is that when someone goes to see a violent movie they lose a few hours of opportunities to be violent. This time-use effect dominates the arousal effect frequently found in lab studies, lending to a re-interpretation of the original research findings. Again, highlighting the great complementarities inherent in conducting lab and field experiments.

In sum, the totality of the evidence causes us to be quite skeptical of Camerer's other major claim:

> Claim: There is no replicated evidence that experimental economics lab data fail to generalize to central empirical features of field data (when the lab features are deliberately closely matched to the field features).

Just considering the evidence within our first sub-section of this section: List (2006a), Bandiera et al. (2005), Benz and Meier (2008), Alpizar et al. (2008) all find that scrutiny is an important determinant of pro-social behavior. The broad hypotheses of each study were replicated in the central features of each data set. This caused the original lab research to not closely match the field data—from the lab, the power of social preferences would be overestimated. This is not denying the existence of such preferences, rather their import in economic settings in the field. Levitt and List (2007b) discuss further evidence. Of course, considering the evidence across the other areas discussed above reinforces the rebuttal of this claim for those areas.

In concluding, while we disagree with all of the bold claims of Camerer (2011), we agree with him fully on one count: more work needs to be done that connects outcomes in the field and the lab. There are only a handful of papers that build a bridge between the lab and the field using AFEs, FFEs, and NFEs, which is really the gold-standard in mediating this discussion (see, e.g., List 2004, 2006a, 2009, and the papers cited above). Nonetheless, we have presented strong reasons to discount Camerer's reading of List (2006a), highlighted a wealth of field experiments that confirm the central contribution of List (2006a) to this debate: it can be difficult to appreciate how findings in the lab relate to observed behavior in natural settings without going to the natural settings themselves.

5. 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 has 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. Fisher and Neyman in the 1920s and 1930s was seminal in that it helped to answer important economic question 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's attempts to evaluate programs by deliberate variations in agency 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, this third period has produced field experiments in economics that have (1) measured 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) been 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 they estimate and the question they can address.

We believe that at this point the field can move beyond strong statements that *lab or field* results will always 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 Volume moves researchers to use AFEs, FFEs, and NFEs to bridge insights gained from the lab with those gained from modeling naturally-occurring data.

References

Al-Ubaydli, O. & List, J.A. (2012). On the generalizability of experimental results in economics. NBER working paper series (no. 17957).

Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Does context matter more for hypothetical than for actual contributions? Evidence from a natural field experiment. Experimental Economics, 11(3), 299-314.

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

Bandiera, O., Barankay, I., & Rasul I. (2005). Social preferences and the response to incentives: Evidence from personnel data. The Quarterly Journal of Economics, 120(3), 917-962.

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

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field? - Evidence from donations. Experimental Economics, 11(3), 268-281.

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.

Budescu, D.V. & Rapoport, A. (1992). Generation of Random Series in Two-Person Strictly Competitive Games. Journal of Experimental Psychology, 121, 352–363.

Camerer, C. (2003). Behavioral Game Theory: Experiments on Strategic Interaction, Princeton: Princeton University Press.

Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. Working paper. Retrieved from Social Science Research Network.

Camerer, C.F., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies, ed. Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (p.55-95). Oxford: Oxford University Press

Camerer, C.F., & Hogarth, R.M. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. Journal of Risk and Uncertainty, 19(1), 7-42.

Camerer, C.F. & Thaler, R.H. (1995). Anomalies: Ultimatums, Dictators and Manners. The Journal of Economic Perspectives, 9(2): 209-219.

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

Cartwright, N. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. History of Political Economy, 23(1), 143-155.

Chamberlin, E.H. (1948). An Experimental Imperfect Market. Journal of Political Economy, 56(2), 95-108.

Dahl, G., & DellaVigna, S. (2009). Does movie violence increase violent crime? The Quarterly Journal of Economics, 124(2), 677-734.

DellaVigna, S., List, J.A., & Malmendier, U. (2012) Testing for Altruism and Social Pressure in Charitable Giving. The Quarterly Journal of Economics, 127(1): 1-56.

Eckel, C.C. & Grossman, P.J. (1996). Altruism in Anonymous Dictator Games. Games and Economic Behavior, 16: 181-191.

Eckel, C.C. & Grossman, P.J. (2008). Subsidizing charitable contributions: a natural field experiment comparing matching and rebate subsidies. Experimental Economics, 11(3): 234-252.

Falk, J. J. , & Heckman, A. (2009). Lab experiments are a major source of knowledge in the social sciences. Science, 326(5952), 535-538.

Fehr, E., & List, J.A. (2004). The hidden costs and returns of incentives - trust and trustworthiness among CEOs. Journal of the European Economic Association, 2(5), 743-771.

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

Fehr, E., Kirchsteiger, G., Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. The Quarterly Journal of Economics, 108(2), 437-459.

Fréchette, G.R. (2012). Laboratory experiments: professionals versus students. Working paper. Retrieved from https://files.nyu.edu/gf35/public/print/Frechette_2009b.pdf.

Gneezy, U., & List, J.A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. Econometrica, 74(5), 1365-1384.

Guatier, P.A., Van Der Kaauw, B. (2010). Selection in a field experiment with voluntary participation. Journal of Applied Econometrics, 27(1): 63-84.

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

Harrison, G.W., List, J.A., & Towe, C. (2007). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. Econometrica, 75(2), 433-458.

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.

Hertwig, R. & Ortmann, A. (2001). Experimental practices in economics: A challenge for psychologists? Behavioral and Brain Sciences, 24(3), 383-403.

Holt, C.A. (1995) Industrial Organization: A Survey of Laboratory Research. in Kagel, J. and A.E. Roth, The Handbook of Experimental Economics, Princeton, (1995), pp. 349-435.

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

Hunter, J. (2001). The Desperate Need for Replications. Journal of Consumer Research 28(1), 149-158.

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

Kessler, J., & Vesterlund, L. (2011). The external validity of laboratory experiments: Qualitative rather than quantitative effects. Working paper. Retrieved from http://www.pitt.edu/~vester/External_Validity.pdf

Knight, F. H. (1921). Risk, uncertainty, and profit. (p. 313). New York, NY: Cosimo.

Lazear, E., Malmendier, U., & Weber, R. (2012). Sorting in experiments with application to social preferences. American Economic Journal: Applied Economics, 4(1):136-163.Levitt, S.D., & List, J.A. (2007a). Viewpoint: On the generalizability of lab behaviour to the field. Canadian Journal of Economics, 40(2), 347-370.

Levitt, S.D., & List, J.A. (2007b). 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. List. (2008). Homo Economicus Evolves. Science 319(5865): 909-910.

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.

Levitt, S.D., List, J.A., Reiley, D. (2010). What Happens in the Field Stays in the Field: Professionals Do Not Play Minimax in Laboratory Experiments. Econometrica, 78(4), 1413-1434.

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, The B.E. Journal of Economic Analysis & Policy, 6(2 - Advances), Article 8.

List, J.A. (2008). Introduction to field experiments in economics with applications to the economics of charity. Experimental Economics, 11(3), 203-212.

List, J.A. (2008), Informed Consent in Social Science," Science**,** (2008)**,** 322(5902), p. 672.

List, J.A., (2009), The Economics of Open Air Markets, NBER working paper series (no. 15420).

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

List, J.A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. The Journal of Economic Perspectives, 25(3), 3-16.

List, J.A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. Experimental Economics, 14 (4), 439–457.

Maniads, Z., Tufano, F., & List, J.A. (2013). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects . Forthcoming, The American Economic Review.

Moonesinghe, R., Khoury, M.J., & Janssens, A.C.J.W. (2007). Most Published Research Findings Are False—But a Little Replication Goes a Long Way. PLoS Med, 4(2), 218-221.

Niederle, M. & L. Vesterlund. (2007). Do Women Shy Away from Competition? Do Men Compete too Much? The Quarterly Journal of Economics, 122(3), 1067-1101.

Palacios-Huerta, I. & O. Volij. (2008). Experientia Docet: Professionals Play Minimax in Laboratory Experiments. Econometrica, 76(1), 71-115.

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). Opting- In: Participation Biases in the Lab. IZA discussion paper series (no. 6865).

Smith, V.L. (1962). An Experimental Study of Competitive Market Behavior. Journal of Political Economy, 70(2), 111-137.

Smith, V.L. & Walker, J.M. (1993). Monetary Rewards and Decisions Cost in Experimental Economics. Economic Inquiry, 31(2), 245-261

Sonneman, U., Camerer, C.F., Fox, C.R. & Langer, T. (2013). How psychological framing affects economic market prices in the lab and field, forthcoming in Proceedings of the National Academy of Science.Stoop, J., Noussair, C.N., & van Soest, D. (2012). From the Lab to the Field: Cooperation Among Fishermen. Journal of Political Economy, 120(6): 1027-1056.

Wacholder, S., Chanock S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. Journal of the National Cancer Institute, 96(6), 434-42.

Winking, J. & Mizer, N. (2013). Natural-Field Dictator Game Shows No Altruistic Giving. Evolution and Human Behavior, in press.

Yoeli, Erez, Moshe Hoffman, David G. Rand, Martin A. Nowak (2013), Powering Up with Indirect Reciprocity in a Large-Scale Field Experiment," forthcoming in Proceedings of the National Academy of Science.