The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List*

Colin F. Camerer

December 30, 2011

Abstract

This paper addresses a recent criticism of experimental economics, by Levitt and List [2007ab, 2008], that lab experimental findings may not generalize to field settings. We argue against the criticism in three ways. First, experimental economics seeks to establish a general theory linking economic factors, such as incentives, rules, and norms, to behavior. Hence, generalizability from the lab to the field is not a primary concern in a typical experiment. Second, the experimental features that could threaten lab-field generalizability are not essential for all lab experiments (except for obtrusiveness, because of human subjects protection). And even so, there is little evidence that typical lab features not necessarily undermine generalizability. Third, we review economics experiments that are specifically designed to test lab-field generalizability; most experiments demonstrated that laboratory findings could indeed be generalized to comparable field settings.

*This paper was prepared for an NYU Methods conference and edited book. Citations from many ESA members were extremely helpful. Support from HFSP and the Gordon and Betty Moore Foundation are gratefully acknowledged. Thanks to Dan Knoepfle, Stephanie Wang, Alec Smith, and especially Rahul Bhui for research assistance, to many patient listeners, to Tim Salmon and Abigail Barr for candid feedback, and to John List for his JPE [2006] data.

1 Introduction

Lab experimentation took serious hold in economics much later than in its neighboring social sciences and in biology. This late development was associated with slow recognition of the possibility and nature of economic experiments. Only twenty years ago, the popular economics text by Samuelson and Nordhaus [1985: 8] stated:

> One possible way of figuring out economic laws ... is by controlled experiments ... Economists [unfortunately] ... cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe.

When it was published in 1985, this passage would have surprised and disappointed the hundreds of economists who actually *had already* done controlled experiments, and published their results in leading journals.

The belated development of experimental economics also seems to be associated with a persistent confusion about the styles of psychology and economics experiments. The blurring of this methodological boundary between psychology and economics may account for some of the ongoing skepticism about generalizability of economics experiments. For example, Gary Becker was quoted in a casual magazine interview as saying:

> One can get excellent suggestions from experiments, but economic theory is not about how people act in experiments, but how they act in markets. And those are very different things. It is similar to asking people why they do things. That may be useful to get suggestions, but it is not a test of the theory. The theory is not about how people answer questions. It is a theory about how people actually choose in market situations. [Clement, 2002]

Becker says that "how people act in experiments" is similar to"asking people why they do things". Lumping the two types of data together is distressing to experimental economics

because our approach is dedicated precisely to creating features of experimental economics which is distinct from surveys (and from many features of experiments in psychology). Indeed, the great methodological achievement of experimental economics has been the creation of experimental and markets which are designed to have the key features– institutional rules, endowments, and incentives– that are considered essential to make predictions from economic theory about behavior (e.g., Smith [1982]).

This paper addresses a recent, vocal criticism about experimental economics: Lab experimental findings may not generalize to field settings. Levitt and List [2007a] expressed this view in an especially provocative and ominous way:

> Yet unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely converge across the lab and many field settings. [LL, 2007a: 364]

Interestingly, a roughly similar debate about generalizability is ongoing in development economics. The debate is about the value of randomized field experiments compared to inferences from larger observational samples. In this debate the imperfections of field experiments are sometimes shown in a harsh light, rather than as a flattering contrast to lab experiments. For example, in a news interview Angus Deaton said

> "They [field experiments in developing countries] tend to be very good at proving causality but often in a very narrow framework. Just because something works in this crazy, local way doesn't necessarily mean that is very useful." [Stanley, 2011].

Defending these experiments, Banerjee and Duflo suggest that

> ... because of an imperfect recognition of what is exciting about the experimental agenda, there is a tendency to set up false oppositions between experimental work and other forms of research. [Banerjee and Duflo, 2009: 152].

This paper has three main points:

First, special concern about generalizability of lab results might result from an aversion to the stylized perspective on what economics experiments are meant to do, that most experimentalists hold (which I call the scientific view).

The *policy view* is that generalizability is crucial, because an implicit goal of lab experiments is to extrapolate from the lab to a particular field setting (or to some imagined setting which is the target of "external validity"). The *scientific view* is that all empirical studies contribute evidence about the *general* way in which agents characteristics, incentives, rules, information, endowments and payoff structure influence economic behavior. This general behavior function is assumed to be parallel in the lab and field. In this view, since the goal is to understand general principles, whether the "lab generalizes to the field" (sometimes called "external validity" of an experiment) is distracting, difficult to know (since there is no single "external" target field setting) and is no more useful than asking whether "the field generalizes to the lab".

Second, even in the scientific perspective it is certainly true that *some* economics experiments have features that make them less likely to generalize to *some* naturally-occurring settings. However, the exact same thing is true of *some* field settings that will not generalize well to other field settings. That is, as the opening quote expresses, concerns about generalizability from lab to field settings should apply to all types of generalization (including field-field).

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– 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. They do not.

Third, we review economics experiments which are specifically designed to generalize from the lab to the field. Only a few experiments were deliberately designed to have properties of field settings and were conducted in parallel naturally-occurring field environments. List's [2006] study of reciprocal quality delivery in sports card markets is looked at in detail. This extra scrutiny is justified because this is the earliest carefully-designed study that compares lab and field differences feature-by-feature (and was prominently published in JPE). In that study, there is no general lab-field difference in reciprocity for all card dealers. There *is* a difference for a minority sample of nonlocal dealers, but it is statistically unreliable (based on new analyses not previously reported). Therefore that study does *not* show especially conclusive evidence of poor lab generalizability. A meticulous study of Dutch fishermen [Stoop, Noussair, and van Soest, 2010] does show more prosociality in the lab than in a comparable field setting. That is the *only* study which shows reliable evidence of poor generalization when lab and field features are closely matched. Furthermore, there are many experiments comparing lab behavior with one group of subjects with similar field behavior from different sample of participants (with no effort to match the lab and field closely). In this type of comparison, there are more than 20 examples of good comparability (particularly for social preference expression) and only two examples with poor comparability.

Finally, note that the phrase "promise of generalizability" in my title describes two kinds of promise. The first "promise" is whether every lab experiment actually promises to generalize to a particular field setting. The answer is "No": The typical experiment promises only to deliver interesting data about the general mapping from variables to behavior, with extraordinary control over independent variables, not to generalize to any particular target setting. The second kind of "promise" is whether generalizability is likely to be "promising" (i.e., typically accurate) for those lab experiments that are specifically designed to have features much like those in closely parallel settings. The answer is "Yes".

2 Is generalizability fundamental? The scientific view

Experimental economists have converged on a reasonably consensual view of what role experimental data play in economics, which minimizes the importance of narrow worries about external validity. In this view, all empirical studies are designed to contribute evidence about the *general* way in which individual characteristics, incentives, rules and endowments influence economic behavior. Experiments contribute especially diagostic evidence by virtue of extraordinary control over independent variables (internal validity). The price of this control is a necessary sacrifice in obtrusive measurement. There is an offsetting bonus of near-perfect replicability.

The guiding idea here is what Vernon Smith [1982] called "parallelism": It is the hopeful assumption that the same general laws apply in all settings. To be sure, the "general laws" can certainly include how behavior depends on a wide variety of parameter values that often covary with lab and field features, but are potentially controllable or measurable as nuisance variables. For example, parallelism does *not* require that students in a lab setting designed to resemble foreign exchange traders behave in the same way as professional foreign exchange traders behave on trading floors. Behavior may differ because the subject populations are different, or because of many differences in the lab and the trading floor. The maintained assumption of parallelism simply asserts that if those differences could be held constant (or controlled for econometrically), behavior in the lab and the trading floor would be the same. Put differently, if many experimental and field data sets were combined, with sufficient variation among variables like stakes, experience, and subject characteristics, a "Lab" dummy variable would not be significant (assuming it did not plausibly correlate with omitted variables). In this very specific econometric sense, the maintained hypothesis is that there is absolutely nothing special about the influence of lab measurement on behavior.

My view is expressed more compactly and eloquently by Falk and Heckman [2009]. They wrote:

> The issue of realism, however, is not a distinctive feature of lab versus field data. The real issue is determining the best way to isolate the causal effect of interest. [Falk and Heckman, 2009: 536]

They express concern that

> The casual reader may mistakenly interpret arguments about realism as an effective critique against the lab, potentially discouraging lab experimentation and slowing down the production of knowledge in economics and other social sciences. [Falk and Heckman, 2009: 536]

The idea that concerns about generalizability are not special to lab experiments is, in fact, found in Levitt and List [2007a] too, at the end of their paper:

> The sharp dichotomy sometimes drawn between lab experiments and data generated in natural settings is a false one. The same concerns arise in both settings regarding the interpretation of estimates and their generalizability outside of the immediate application, circumstances, and treated population.

Given the consensus among most experimental economists that realism, generalizability, or external validity are not especially important, it is useful to ask where that concern came from, historically. The term "external validity" was coined in the short landmark book on experimental design by Campbell and Stanley [1963] (cf. Bracht and Glass [1968]). The book was originally published as a chapter in the *Handbook of Research on Teaching*. Stanley was an educational psychologist. His interest in external validity probably originated in thinking about educational experiments– e.g., what teaching methods produce the best learning under highly specific ecological conditions. Along these lines, Campbell and Stanley [1963: 5] note that

> While *internal validity* is the *sine qua non* [essential condition] and while the question of *external validity*, like the questions of inductive inference, is never

> completely answerable, the selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of *known character* is the desideratum. (*italics* mine)

An interest in "generalization to applied settings of known character" is therefore required, in order to care about and judge external validity. But this interest in applied settings of known character is the exception in experimental economics: Experimental tests 'generalize to' the general behavior function assumed (by parallelism) to apply everywhere, not to any specific setting. But then there is no basis on which to judge or doubt its external validity. Of course, if the purpose of an experiment is to supply a policy-relevant answer to a particular external setting, then it is certainly important to ask about how well the experimental setting resembles the target external setting to judge its external validity.

These issues are brought into sharp relief by an example: Field data from high-stakes game shows with clear rules. Such data have been usefully exploited in many high-profile studies to learn about risky choice and strategic thinking [Andersen et al., 2008]. But are game shows more "externally valid" than economic choices in canonical lower-stakes lab experiments? Not necessarily. As Bardsley et al. note,

> Such sources of data are often seen as "natural experiments". However, in this case the term "natural" should be interpreted with care. Although game shows are "natural" in the sense of arising independently of the research process, there are other respects, such as the nature of the decision problems and the presence of a presenter and studio audience, in which they seem quite untypical of normal life. Questions can therefore be posed about their broader external validity [Bardsley et al., 2009: 26]

Game show field data *are* very useful, but not for their superior external validity compared to lab experiments. Game shows typically feature simple choices which can be analyzed clearly

to judge whether participants are rational, for higher stakes than in the lab. So game shows are an excellent source of data on simple decision making quality under high stakes...but game show decisions also have extreme audience effects, unusual (often opaque) participant selection, and imperfect control for participant comprehension.

3 Generalizability as discussed by Levitt and List

Let's now turn to a discussion of generalizability as articulated in three papers by Levitt and List [2007a,b, 2008] (LL). One of their papers begins by saying that

> We can think of no question more fundamental to experimental economics than understanding whether, and under what circumstances, laboratory results generalize to naturally occurring environments. [LL, 2007a: 347]

LL "take as given that the issue of generalizability is important". Cited for support are the psychologists Pruitt and Kimmel [1977] and Anatol Rapoport [1970]. Notice that all of those references are about 1970s experimental psychology, not about modern experimental economics. They also note a more subtle critique by Bardsley [2005].

Arguing *against* the importance of generalizability, they cite experimental economists Charles Plott, Vernon Smith, and Arthur Schram ("the 'classic' view of experimentation is one of theory testing, arguing that 'external validity is not an issue' in these instances." [LL, 2007a: 349]).

The "classic" view is sometimes called "blame the theory" by Bardsley et al. [2009]. The idea is this: Since the theory as enunciated does not say that students in a lab behave differently than NYSE floor traders, for example, then an experiment which ignores that distinction is a valid test of a theory which also ignores the distinction. The phrase means that "the theory" is blamed for not including a role for student-trader differences. I do not like the phrase "blame the theory" because a person (or institution) must be blamed

in order to make progress. The burden of hypothesizing (not "blame", per se) should rest somewhere. It would be useful for the economics profession to develop guidelines on where the hypothesizing burden should rest. Put concretely, if a critic says "Your experiment is not externally valid", that criticism should include content about what external validity is expected, why (is it written into the theory? should it be?) and what new experiments could be run to satisfy the critic's concern.

Comparing methods for empirical inference

Following Falk and Heckman [2009], suppose that an outcome of interest Y is fully determined by a function $Y = f(X_1, X_2, ..., X_N)$ (an "all-causes" model, excluding pure error). Suppose we can establish causality in a lab or natural experiment by controlling one variable $X1$ and holding the other variables fixed or measuring them (denote them X^*). This vector X^* includes agent characteristics, endowments, incentives, rules, context, knowledge of other agents, and payoff structure. In lab experiment a vector X^* commonly– but not necessarily – consists of student subjects making abstract choices earning modest payoffs, who are carefully instructed about the relation between their choices and their payoffs. A common critique of the results of lab experiments is that the causal effect of $X1$ in the presence of lab X^* is different than behavior

> in a 'natural setting', that is, for another set of conditions X^{**} including, for example, different institutional details, payoffs, and a different participant population. [FH, 2009: 536]

Falk and Heckman specifically note the apparent difference when X^{**} is sports card dealers [List, 2006] compared to lab experiments on reciprocity in offering higher product quality in response to higher prices. They note that

> If one is interested in the effect of social preference under a third condition X^{***} neither the undergraduate [lab] nor the sports-cards field study may identify the effect of interest. It is not obvious whether the lab X^* or the field X^{**} is more

> informative for the third condition unless a more tightly specified econometric model is postulated or a more precisely formulated policy problem is specified. [FH, 2009: 536]

That is, it is simply not clear that the lab conditions X^* or the card dealer conditions X^{**} are more like the "real world" if the particular target world is some X^{***}. This is the central issue: If the litmus test of "external validity" is accurate extrapolation to X^{***}, is the lab X^* necessarily less externally valid than the field setting X^{**}? How should this even be judged?

Since judging external validity is so slippery, let's return focus to control. There is little doubt that the quality of control is potentially very high in lab experiments (and measures of control quality– internal validity– are available). The usual analysis of field data typically has no direct control. Instead, econometric techniques and judgment are used to infer causality.

Ideal studies with field data exploit natural experiments in which there is either true random assignment of agents to treatment groups, or a good instrumental variable that mimicks natural random assignment (see Angrist and Krueger [2001]) to achieve control. However, the most convincing field studies with truly exogeneous instruments do not necessarily come from domains which are particularly "generalizable". Instead, wonderful instruments are often created by unusual quirks in policy or accidents of history. Their generalizability to other times and places is therefore in clear doubt (if skepticism about lab-field generalizability is applied uniformly). These doubts are accepted, as they should be, as a price to pay for unusually good natural-experimental control.

Note well that this logical claim is not at all meant to impugn the value of the most extraordinary quasi-experimental field studies. Instead, the point is just that the "best" field studies for establishing causality can be suspected of being difficult to generalize just as the "best" method (experimental control) is suspected of the same weakness.[1]

[1]There is also a lively debate about the value of instrumental variable IV studies versus other techniques (e.g., Heckman and Urzua [2009]; Imbens [2009]; Deaton [2009]).

Which methods most rapidly build up knowledge about $Y = f(X_1, X_2, ..., X_N)$? Each method has advantages and they have dynamic complementarities. Discoveries using one method are often then best explored further with another method. In the scientific view, the best empirical studies of all kinds (lab and field) seize on unusually diagnostic opportunities to measure $Y = f(X_1, X_2, ..., X_N)$ or to compare competing theories.

In comparing methods, lab experiments do have three special features that distinguish them: Near-perfect replicability, measures of cognition, and obtrusiveness.

Because instructions, software, recruiting protocols, databases, and statistical tools are so easy to archive and reproduce, a lab experiment can be *replicated* very closely across a huge span of times and places. (And even the threat of replicability makes most experimental economists very careful.) In the lab it is also possible to measure many aspects of *cognition* and neural behavior relatively easily. However, lab experiments are *obtrusive*: Because subjects must give informed consent (in the US and many other countries), they know that they are in an experiment. Below we discuss further whether knowing you are being 'experimented upon' has strong reliable effect on economic behavior that is different than comparable effects in parallel field settings. I think it has little impact except in a few narrow settings.

How do field experiments compare to lab experiments on these important dimensions of replicability, measuring cognition, and obtrusiveness?

Replicability is typically lower for field experiments than for lab experiments[2], and could be impossible. When access to a field site is controlled by a company, government or NGO, new researchers could typically not replicate precisely what previous researchers have done.[3] And since field settings are naturally-occurring, they sometimes change substantially over time (remote villages develop, firms go bankrupt, etc.) which undermines ideal replicability. The fact that they are *not artificial*– and hence cannot be recreated over and over– then becomes an *impediment* to ideal replicability.

[2]Uri Gneezy made this point in a panel discussion at a Berkeley SEED-EC conference, December 2, 2011.
[3]Banerjee and Duflo [2009].

Field settings are also not typically ideal for measuring fine-grained cognition at the individual level (though it is often possible, in principle).

Unobtrusiveness can be extremely good in naturalistic field experiments (such as List [2006]) if the subjects do not have to give informed consent (if their behavior is publicly observable). However, many field experiments are conducted in collaboration with a government or NGO (typically so, in development). In these settings, subjects may sense if there is a change in economic variables (such as a new subsidy) or that they are being experimented upon; they may also have a strong desire to please experimenters to get future opportunities.[4] If the worry about experimental obtrusiveness is that it produces special changes in behavior, it is possible that feeling "experimented upon" by a Western NGO in a poor country, for example, has a much stronger effect on behavior than being experimented on by a professor in a college lab.

4 Potential threats to generalizability

The notation $Y = f(X_1, X_2, ..., X_N)$ implies that whether a conclusion generalizes from one method and setting to another setting is the same as asking how different the $X_1, X_2, ..., X_N$ variables are in the two settings, and how sensitive the $f(.)$ function is to those differences. Conclusions are likely to generalize if all features of two settings are close or if $f(.)$ is insensitive to the features that differ. Note well that lab-field differences play no special role in this calculus of generalizability, except if experimental and domain features of X^* inexorably differ (the "alteration" Bardsley et al. [2009] refer to).

It is true that in the past, many lab experiments in economics have tended to use a common configuration of design features. Typically behavior is observed obtrusively, decisions are described abstractly, subjects are self-selected volunteers from convenience samples (e.g., college students), and per-hour financial incentives are modest. Let's refer to this type of

[4] Chris Udry made this point in a panel discussion at a Berkeley SEED-EC conference, December 2, 2011.

experiment as the "common design".

It is crucial to note that the common design is not the only design. *All* of the common design features have been measured and varied tremendously in lab experiments.[5] If the goal is to extrapolate from conclusions to field settings that *do not* share most of these features, it is worthwhile to consider the threats to generalizability from the common design to such field settings. Doing so, Levitt and List [2007b] suggest that

> behavior in the lab is influenced not just by monetary calculations, but also by at least five other factors: 1. The presence of moral and ethical considerations; 2. the nature and extent of scrutiny of one's actions by others; 3. the context in which the decision is embedded; 4. self-selection of the individuals making the decisions; and 5. the stakes of the game. [LL, 2007b: 154]

Note that the above sentence is equally persuasive if the word "lab" is replaced by "field". So the issue is whether factors 1-5 matter for lab behavior in a way that is never parallel to their influence in the field. Next we will examine the arguments and evidence about whether these features necessarily vary in the lab and field, and how sensitive behavior is to these features in their special lab incarnation.

A. Presence of moral or ethical considerations

There is no doubt that moral and ethical concerns trade off with self-interest, in the lab and in the field. The issue for lab-field generalizability is whether lab settings provide systematically misguided evidence about the effect of moral and ethical concerns in comparable field settings. To be concrete, the issue is whether an experimental subject donating money in an artificial lab environment, knowing experimenters will see what they donated, necessarily creates moral and ethical forces which are fundamentally different than those at work for example, when Sunday parishioners put money in a donation basket that is being

[5] On subject pool effects, see Ball and Cech [1996], Frechette [in press], and Henrich et al. [2010].

passed around while other parishioners watch and donate, knowing priests will add up all the donations later.

There is actually little evidence that experiments are misleading compared to field settings with similar features, and there is much evidence that lab-field results are close (see section 5 below). It is certainly true that in experimental dictator game allocations, when subjects are asked how much of $10 unearned income they would give to a stranger recipient, most give nothing; but a few give $5 and some give $1-2. The overall rate of giving is about 15% [Camerer, 2003: Ch. 2]. A benchmark for giving in the field is the rate of typical charitable contribution as a fraction of income, which are typically around 1%, so it is true that experimental subjects in many dictator games do give a larger fraction of *unearned* income to *strangers*, when prompted, than the fraction of *earned* income people typically give in field settings to known *charitable causes*.

Keep in mind that the dictator game was not specifically designed to predict everyday sharing from earned income. Seen this way, the difference in giving in lab dictator games and in everyday charities is no surprise. However, commentors often describe the dictator lab results as if predicting everyday giving was the goal of the experiment. For example, Levitt and Dubner [2009] quote John List as writing

> 'What is puzzling, he wrote, is that neither I nor any of my family or friends (or their families and friends) have ever received an anonymous envelope stuffed with cash.

It is true that many early accounts of the scientific role of the dictator game called it a way to measure "altruism", in contrast to other games that are thought to express combinations of reciprocity and other forces (e.g., Camerer [2003]). (A better phrase would have been "nonreciprocal sharing", or somesuch.) Evidence accumulated since then suggests that the dictator game does not measure pure altruism, per se, but instead measures a willingness to conform to a perceived social norm of appropriate sharing of unearned income.[6] This social

[6]See also Andreoni and Bernheim [2009]; Krupka and Weber [2008]; Bardsley [2005]; List [2007].

image motive is expressed by the term "manners", first used by Camerer and Thaler [1995].

The social image or norm interpretation implies that allocations *should* vary with any details of procedures and possible allocations that affect image and norms. Indeed, the extreme control in the lab suggests it is an ideal setting in which to learn about influences on sharing. The nature of entitlements, deservingness, stakes, and obtrusiveness (how much you are being watched, by whom) can all be controlled much more carefully than in most field settings.

B. The nature *and* extent of scrutiny of one's actions by others

A lab experiment is "obtrusive" (to use a conventional experimental term, see Webb et al. [1999]).[7] That is, lab subjects know their behavior is being recorded and will be studied scientifically. The question is whether this obtrusiveness changes behavior in a way that is different than in a field setting with comparable obtrusiveness.[8]

The danger here is that being obtrusively watched by a lab experimenter creates demand effects which are unique to the lab, and can never be controlled for if they were identified (an "alteration" problem, see Bardsley [2005] and Bardsley et al. [2009]). To an economist, a demand effect is the result of a perception by subjects of what hypothesis the experimenter prefers, along with an intrinsic incentive subjects have to cooperate by conforming to the perceived hypothesized behavior. In this view the subject is an eager helper who tries to figure out what the experimenter) wants her to do.

First note that there is no clear evidence of experiment-specific demand effects of this type in modern experimental economics; there is merely suspicion.[9]

[7]Others (e.g. LL [2008]) use the term "scrutiny" which will be treated as synonymous with the more established term "obstrusiveness" in this paper.

[8]Many field experiments, in development for example, are clearly as obtrusive as in university labs and may create even stronger potential experimenter effects if the subject-experiment relationship is less familiar to low-education rural villagers, say, than it is to college students.

[9]An important example is the double-blind dictator experiments [Hoffman et al., 1998]. One possible interpretation of those results is that subjects in the standard game give because they believe that the experimenter wants them to. (Note that in this design self-interested behavior also occurs at a boundary of the choice set, so any random deviation looks like both prosocial giving and like expression of satisfying a perceived demand to give.) The drop in giving in the double-blind case could be consistent with disappearance of such a perceived demand to give. But it could also be consistent with changing the subject's belief that

Why might strong reliable demand effects be rare in economics experiments? For strong demand effects to exist and to jeopardize generalizability, subjects must

(a) have a view of what hypothesis the experimenter favors (or "demands"); and (b) be willing to sacrifice money to help prove the experimenter's hypothesis (as they perceive it).

In most economics experiments involving choices, games or markets, condition (a) is just is not likely to hold because subjects have no consistent view about what the experimenter expects or favors. Instructions are also very carefully written to avoid instilling any such expectations.[10]

Furthermore, in many economics experiments there are two or more competing hypotheses and experimenters themselves are not sure which hypothesis will be true. (This is particularly the norm in behavioral economics experiments in which a rational hypothesis is often pitted against a behavioral alternative.) In these cases, even if the subjects did have accurate expectations of the range of what the experimenters expect, it would not tell them what to do. As Falk and Heckman [2009] say

> It [obtrusiveness] is a minor problem in many experiments, especially if the decision environment is interactive and rich, such as in sequential bargaining or market experiments. Moreover, being observed is not an exclusive feature of the laboratory: many decisions outside the lab are observed. [p. 537]

Some evidence about condition (a), subjects' expectations of experimental demand, comes from Lambdin and Shaffer [2009]. They replicated three classic experiments showing different types of preference reversal. Subjects chose between two outcomes in each of two treatment conditions. In each treatment condition the outcome consequences are the same, but the descriptions or choice procedures differ. (One example is the "Asian disease" problem showing gain-loss framing effect). Because the choices have identical consequences in

their specific allocation will be known to anybody, not just to an experimenter.

[10]Experimental economists work hard to eliminate demand effects. Orne [1962] advocates trying to measure demand effects in various ways (e.g., presenting passive subjects with the experimental materials and asking them what they think is likely to happen or what the experimenter expects). A modern study using these methods would be useful.

the two conditions, it is possible that subjects might easily guess that the hypotheses have to do with ways in which choices would actually differs or not differs.

One group of subjects were shown both of the treatment conditions and asked whether they could identify the experimental research hypothesis. Subjects were quite confident that they could guess (mean 3.35 on a 7-point scale, with 1 as "extremely confident"). Eighty percent of the subjects said the experimental hypotheses were "completely" or "somewhat" transparent. The authors also asked members of the SJDM society (who study questions like these); most of them (71%) thought the hypotheses were transparent too.

However, the subjects' guesses about research hypotheses were only accurate 7%, 32% and 3% of the time (across the three examples).[11] Thus, even very simple experiments that are the most likely to create demand effects spuriously producing experimental results, do not necessarily do so, violating condition (a) for demand effects.

For the sake of argument, assume that condition (a) does hold, and subjects accurately know that experimenters expect and favor a particular outcome. The next necessary step for demand effects to matter is condition (b), that the subjects will sacrifice earnings by behaving in a way the experimenter favors. Even if there is a desire to satisfy the experimenter's perceived demand (condition (b)), if that desire is simply a component of overall preference then clear financial incentives should be able to overwhelm any such preference. A natural test for existence of a hypothesized demand effect is therefore to see if apparent demand effects shrink when more money is at stake. The fact that increasing stakes from modest levels to much higher levels typically has little effect (e.g., Camerer and Hogarth [1999]) suggests that as long as there is some salient incentive, demand effects are not playing a large role.

As a matter of logic and empirics, the case for clear, substantial demand effects in economics experiments is weak: Subjects are not necessarily likely to guess what economics

[11] They also included a truly transparent new experiment (whether a person was more likely to buy a $10 school raffle ticket if they were given a free can of soda), and 88% of the subjects did correctly guess the experimenters' hypothesis.

experimenters favor (because the experimenter favors no particular result, or because what they favor is opaque); and even if subject do guess what experimenters favor, they may not sacrifice much in earnings to help the experimenter; and if they sacrifice earnings to help the experimenter, we can remove those effects (and estimate their elasticity) by paying more.

So where does the belief that strong demand effects exist in modern economics experiments come from? My hunch is that it comes from confusion between the canonical older methods of experimental psychology and modern methods of experimental economics. Those two systems of methods developed largely independently and are fundamentally different.

Some support for the confusion hypothesis comes from the historical sourcing in LL's writing. They assert that:

> Decades of research within psychology highlights the importance of the role obligations of being an experimental subject, the power of the experimenter herself, and the significance of the experimental situation. [LL, 2007b: 158]

To illustrate the dangers of demand effects in modern experimental economics, LL quote from an obscure short rejoinder made 100 years ago by A. H. Pierce [1908] about how cooperative experimental subjects are. Pierce's comments were part of a debate about whether hysterical patients tell psychiatrists what they think the psychiatrists want to hear. LL also cite a classic fifty-year old paper by Orne [1962]. Orne's discussion of demand effects was also partly motivated by his particular fear that hypnotized patients were overly cooperative (i.e., they respond to post-hypnotic suggestions in order to "help out"). The psychological and psychiatric sources LL cite are interesting for distant historical perspective, but possible demand effects in studies of hypnosis and hysteria 50 and 100 years ago have nothing to do with modern economics.

LL also write:

> Schultz [1969: 221] described the lab as having a superior-subordinate relationship matched only by that of parent and child, physician and patient, or drill sergeant and trainee. [LL 2007b: 159]

Readers who have either been subjects or experimenters should judge for themselves how farfetched Schultz's comparisons are.[12] If experimenters could guide the subjects' behavior so strongly, and chose to, it would be a lot easier to publish papers!

To illustrate their fears concretely, Levitt and List [2007a] describe the subject's-eye view of (a fictional) "Jane", who participates in an abstract experimental game designed to test theories of firm behavior. Their discussion of obtrusiveness, subject naivete, and demand effects in this tale simply rings false to an experienced lab experimenter. It bears little resemblance to typical subject participation in the active working labs at places like Caltech and many other institutions with long-lived experimental economics research programs.

At one point fictional Jane signs a consent form indicating that she

> understands her decisions are being monitored, recorded, and subsequently scrutinized for scientific purposes. Jane realizes that she is entering a relationship that has no useful parallels in her everyday life. [LL, 2007a: 348]

I disagree : The relationship Jane is entering actually has *many* useful parallels in her everyday life.

First, if Jane has been in lab experiments before, then she *does* have a useful parallel– her previous experimental experience. There are "no useful parallels" only if Jane is an experimental virgin.

Second, even if she is an experimental virgin, she will probably have been in many situations in which she makes abstract choices, given some simple rules, and her choices have economic consequences for her and will be recorded and analyzed by some people or system that created those choices and are interested in what she does. For example, Jane has done homework and taken exams. She took an SAT test in which her decisions are 'recorded and subsequently scrutinized' for college admissions purposes. She probably played card games or board games with friends with abstract rules leading to wins or losses. She may

[12] As a parent I indulge the right to comment on the parent-child analogy. My son has sometimes started to cry instantly, because of an effective disappointed-Dad look; he also sometimes shouts that he "hates" me. Subjects don't do this.

have had job interviews, including written or oral exams, which are scrutinized for hiring purposes. She may have played a sport in which a coach scrutinized her performance, with the hope of being chosen for a team or given an award. All of these activities constitute potentially useful parallels in her everyday life for the experimental relationship.

C. The context in which the decision is embedded

A major achievement of psychology, behavioral economics, and experimental economics (working in a loose consortium) has been to establish that contextual features and cues can have a substantial impact on behavior. Evidence of contextual effects has been building up for four decades, almost entirely from lab experiments in which context can be varied and all other differences can be controlled. LL go a little further and assert that "context matters and is not completely controlled by the experimenter" [LL, 2007b: 163]. Context therefore makes their top five list of threats to generalizability of lab results.

Let's examine their argument closely: (1) LL argue that context matters (mostly) because experiments have conclusively shown that context matters by varying context and controlling other variables. Then (2) they fear that since context "is not completely controlled by the experimenter", lab contexts cannot be equalized effectively to comparable target field contexts.

The irony of their pessimistic conclusion, about lab context generalizing inconclusively to field context, is that premise (1) depends on strong experimental control of contexts, while premise (2) depends on weak experimental control of context. Both cannot be right. If context with good control is achievable in the lab, then it should also be possible to create field-like context, if desired, in the lab too. Furthermore, if context is not *completely* controlled in an experiment, it is not any more controllable (or measurable) in typical field experiments or in naturally-occurring data.

In any case, LL cite two types of data which are supportive of this hypothesis that uncontrolled context matters. Let's discuss one type of data, the artificial cross-cultural field experiments by Henrich et al. [2005] (in which I participated as a coach). They found that

local sharing norms in analogous choices[13], reported informally by the anthropologists, and cross-cultural ranked measures of cooperation and market integration themselves correlate with ultimatum offers. LL note that "the context the actors brought to the game and that experimenters cannot control– like past experiences and internalized social norms– proved centrally important in the outcome of play". It is true that these cultural sharing norms were not controlled by the experimenters, because the naturally-occurring variation in norms was the independent variable of interest. However, variables associated with sharing norms could be *measured* to interpret different behavioral patterns (and correlated with ultimatum offers; see Henrich et al. [2006]). Just as in field experiments and analyses of field data, failing to completely control everything in a lab experiment is only a serious flaw if the uncontrolled independent variables cannot be measured and statistically controlled for.

D. Self-selection of the individuals making the decisions

Potential subjects are usually told a little about an experiment before they volunteer. Then there is self-selection that might make the resulting subject pool different than the recruited population, and maybe different than in many field settings. (These are often called "volunteer effects"). One view about selection, based on old concepts in experimental psychology is that subjects are "scientific do-gooders" who "readily cooperate with the experimenter and seek social approval" [LL, 2007b: 165].

"In contrast", LL write, "market participants are likely to be a selected sample of individuals whose traits allow them to excel in the marketplace".[14] They suggest that self-selected experiment volunteers might be more pro-social than a random sample and less pro-social than "market participants". But Hoffman and Morgan [2011] found the opposite: There is more prosociality among workers in two internet industries (domain name trading and adult entertainment) compared to students. Several other studies have found that, if anything,

[13]After figuring out how a public goods game worked, an Orma participant in Kenya exclaimed "Harrambee!" because the game reminded her of a practice used to raise funds for public goods (such as building a school) in which a village official harangues people to contribute one by one [Ensminger, 2004].

[14]Selection is complicated. Overconfidence, discrimination, and network effects all mitigate the idea that there is pure self-selection of people into markets those people are best suited for.

students are clearly less prosocial than working adults (e.g., Alatas et al. [2009]; Belot, Duch, and Miller [2010]; Anderson et al. [2010] and see Frechette [this volume] on general subject pool differences).

In fact, there is very little evidence that volunteer subjects are systematically different than nonvolunteers on traits or behaviors. LL report [LL, 2007b: 166] that sports-card sellers who volunteered for List [2006] study were not significantly different in reciprocity than nonvolunteers (as measured by their later behavior in an unobtrusive field experiment). Eckel and Grossman [2000] found that volunteers were *less* pro-social. Cleave, Nikiforakis, Slonim [2011] find no volunteer biases in a trust game and lottery choice task among 1173 students. Anderson et al. [2011] found no difference in prosociality among self-selected adults and a comparable sample of adult truck driver trainees who are essentially compelled to participate (91% do so).

A likely reason why there are not systematic volunteering effects is that subjects volunteer to earn money, not because they are "scientific do-gooders" (as LL conjecture). In schools with active economics labs subjects *do* see themselves as "market participants whose traits allow them to excel in the marketplace" *...in the market place for experimental labor.* Students work in experiments as temporary employment, earning money for interesting work with flexible hours.

Finally, since self-selection into market activity *is* an important force in economic behavior, a question arises about how selection can be studied most conclusively. Ironically, while unmeasured selection into experiments is a potential threat to generalizability, controlled lab experiments provide some of the best possible methods for learning about self-selection. The ideal approach to measuring self-selection is to measure characteristics of a large pool of subjects, allow them free choice into different experimental conditions, then see which subjects choose which condition and how they behave (see, e.g., Lazear, Malmendier and Weber [forthcoming]). For example, suppose you wanted to test whether market participants are a self-selected sample of individuals whose traits allow them to excel in the marketplace. An

ideal way to do so is to measure their traits, allow them to select into different marketplaces, and see who selects into what markets and whether they excel. This is typically very difficult in field data and easy in the lab.

E. The stakes of the game

The possible influence of economic incentives ("stakes") on behavior has been explored, many, many times in experimental economics. Smith and Walker [1993] examine about two dozen studies and concludes that there is support for a model in which higher stakes motivate more cognitive effort, reducing response variance and bringing responses closer to rational ones. Camerer and Hogarth [1999] look at a larger sample of 74 studies. They conclude that the largest effects of incentives come from comparing hypothetical payoff studies (with zero marginal financial incentive) to those with some incentives. The most reliable effect is that noisy response variance falls. In some cases (e.g., judgment tasks where there is a normatively correct answer) higher incentive induces more rational choices. Incentives also seem to reduce effects of socially desirable behaviors (e.g., there is less dictator giving, and less risk-taking, with higher incentives).

Most studies in foreign countries where purchasing power is low have found that the basic patterns in a variety of games and markets which are observed at modest stakes in traditional subject pools also replicate when stakes are very high. Therefore, the effect of stakes is no longer a sound basis for critiquing lab-field generalizability. It is also frustrating that the common assertion that raising stakes will change behavior is almost never expressed in the form of how *much* behavior will change in response to a change in stakes. It is true that in some field settings with very high stakes, it may be impossible as a practical matter to conduct lab experiments with matched high stakes. But there is also little evidence that substantial differences in stakes (within ranges that have been varied experimentally) change conclusions derived from modest stakes.[15]

[15]Andersen et al. [in press]'s study of high-stakes ultimatum game bargaining among the Khasi speakers in India does show an apparent decrease in rejections of percentage offers as stakes increase (though typical rejected offer amounts also increase with stakes). Their result actually supports the conclusion from most

F. New experiments are the best remedy for concern about generalizability of older experiments

Falk and Heckman [2009: 537] note that

> Ironically, most objections [concerning lab evidence] raise questions that can be very well analyzed with lab experiments, suggesting the wisdom of conducting more lab experiments, not fewer.

To illustrate Falk and Heckman's point, recall LL's list of candidate factors that limit lab generalizability: Moral considerations, scrutiny, context, self-selection, and stakes [LL, 2007b: 154]. The most conclusive evidence that variation in these factors changes in behavior *comes from the lab itself.* The lab evidence is the most conclusive because field measures (and natural experiments) have much weaker control over those factors than lab environments do.

Above I discussed how lab experimental control might prove especially useful in studying the effect of self-selection (although it has not been used very often so far). Another example is scrutiny (obtrusiveness). The important part of obtrusiveness that we care about is how it is perceived by the economic agent. Measuring obtrusiveness requires us therefore to vary its level and nature, *and* to accurately record how it is perceived by the agent. In naturally-occurring field examples, obtrusiveness might well vary (e.g., supervision of worker effort) but the variation would typically be highly endogeneous from firm choice– hardly exogeneous– and often very difficult to measure. There are beautiful examples of such variation in field experiments (e.g., Rebitzer [1988]), but a high degree of control is hardly ever attained in typical field data without quasi-experimental control.

5 Examples

So what do the best available data about lab-field generalizability say?

modest-stakes ultimatum experiments, which is that rejections express some type of social preference (e.g., trading off monetary share and money, as in Bolton and Ockenfels [2000]) that predicts the observed stakes effect.

I will focus mostly on two kinds of data. The closest lab-field comparisons are carefully designed to have the features of particular field settings, and compare the behavior of the same types of subjects in the field and in the lab facsimile. Unfortunately, there are a small number of these close comparisons (only six). There are many more examples of lab experiments which measure a behavior (such as prosocial cooperation) in one subject pool, and a related field behavior which is not closely matched in structure, but which is expected to be correlated and that individual or group level with the lab results.

A. The closest tests of lab-field generalizability

Sports card trading: The best series of closely-comparable lab and field experiments uses transactions with sports card dealers, conducted over different years and compiled in List [2006].

The setting is sports-paraphernalia dealer conventions. Dealers arrive and spread their goods for sale on tables. Buyers mill around, often request a particular quality level (for cards which can be independently graded) and offer a price. The crucial question is whether dealers offer items of the requested quality, or give lower-quality items.

For brevity, let's focus on the lab experiment "lab-market" which most closely matches the economic environment in three of the field experiments. (Five other treatments are interesting but are not such close matches.)

The transactions are simple: Buyer subjects are recruited (for a flat participation fee) to approach actual dealers and offer either $20 requesting a card with quality 4, or $65 requesting a card with quality 5. Note that these experiments are unusual because the buyers have no financial stake in the outcome.

One key to the success of the study is that the cards which are actually bought by subjects can then be independently graded to compare the "actual" (graded) quality with what was requested. The experiment takes place near the field marketplace and is closely matched in protocol. There are two key variables in the market of interest (given that prices are fixed by experimental control): First, how much quality is supplied by dealers? And second, does

supplied quality respond to price?

The first finding is that there is a clear drop in quality from the lab to the field. The average quality levels for the $20 and $65 cards are 3.1 and 4.1 in the lab, and 2.3 and 3.1 in the field. So the shortfalls in quality from what were requested are around one quality point lower in the field. All dealers seem comfortable delivering less quality in general in their naturally-occurring environment (when they do not know if quality is later measured).

The second finding is that the response of quality to price is closely comparable in the lab and field. Table 1 summarizes some of the numbers in List's Table 4. The lab-context experiment has a price coefficient of .05 but its larger magnitude could be due to the fact that buyers are incentivized for performance in that treatment (and are not incentivized in the others reported).

TABLE 1 HERE

In the other treatments with no buyer incentive the coefficient of quality on price is .02 in every treatment. This implies that paying an extra $50 will increase provided quality by about one unit. The statistical strength of this gift exchange effect is actually *stronger* in the field card market than in the lab markets (probably just due to larger sampling in the field).

Table 1 could hardly show better lab-field generalizability. Effects of price on quality in the lab and the field (the "price coefficient") are exactly the same.

However, in *Science* Levitt and List [2008] wrote:

> Some evidence thus far suggests that behavioral anomalies are less pronounced than was previously observed in the lab [List, 2006; Figure 1]. For example, sports card dealers in a laboratory setting are driven strongly by positive reciprocity, i.e., the seller provides a higher quality of good than is necessary, especially when the buyer offers to pay a generous price. Yet, this same set of sports card traders in a natural field experiment behaves far more selfishly. They provide far lower

> quality on average when faced with the same buyer offers and *increase quality little* in response to a generous offer from the buyer. [List, 2006: 910] (italics mine)

Hmm. This paragraph– and especially the italicized phrase– does accurately describe the full results, as summarized in Table 1. . What is going on? The discrepancy is because the *Science* reference to "sports card traders in a laboratory setting" refers to the minority segment of the sample (30%), nonlocal dealers only, not to the full sample (without making that reference explicit). There is a substantial difference in behavior between "local dealers" (who usually go to the conventions repeatedly and may operate a storefront business or website) and "nonlocal dealers", as self-identified by surveys. Their conclusion refers only to the nonlocal dealers and ignores the larger group of local dealers.

FIGURE 1 HERE

Figures 1a-b shows the average quality for both nonlocal and local card dealers in the lab and field experiments. (Figure 1a is the one reported in LL [2008] with standard error bars approximated). There is apparently a sharp difference between responsiveness of local and nonlocal dealers in the field, but no such difference in the lab.

The local-nonlocal comparison is indeed important, because the reciprocity exhibited by the locals *could* be driven by reputational concerns (for local dealers only), by social preferences, or by both. If the non-locals have lower reputational concerns their behavior does provide better isolation of pure social preference effects.

TABLE 2 HERE

A fresh look at raw the data[16] shows two new observations not reported in the original paper.

The key data for each dealer is the difference between the quality supplied for the $20 card and the $65 card. Define the quality response as the actual quality offered for a $65

[16]Data were generously supplied by John List.

card minus the quality offered for a $20 card. The important comparison is whether that quality difference is larger– more reciprocal– in the lab than in the field.

These results are summarized in Table 2. The non-locals offer 1.11 more quality units in the lab for the more expensive card. The nonlocals only offer .13 more in the field, so they are indeed less reciprocal in the field than in the lab. However, the locals are somewhat *more* reciprocal in the field (1.40 compared to .95). **That difference is the first new observation.**

In List [2006] and Levitt and List [2008], the difference in reciprocity (i.e., the seller's quality response to the buyer's offered price) between lab and field was not tested statistically. The proper test is a difference-in-difference test, in which the quality differences for high and low prices are compared in the different lab and field locations. Because the distributions of quality differences are discrete and not at all Gaussian, nonparametric tests are appropriate. Three different tests give different results [Table 2]. One test (Fligner-Policello) shows a strongly significant negative difference for the nonlocal dealers (less field reciprocity) and a weaker positive effect for locals. Another test shows the opposite conclusion about significance, more strongly (Epps-Singleton). Overall, the conclusion that the lab and field show different reciprocity is just not robustly significant (in two or more of the three tests). **That's the second new observation.**

This conclusion is certainly more equivocal than Levitt and List's [2008] claim. Their claim is numerically correct– the average quality difference of 1.11 in the lab is greater than .13 in the field. However, the test is underpowered and hence not statistically conclusive. List [2006: 32] concludes from this:

> The data suggest that social preferences do not have a major impact in these particular markets.

This conclusion goes in the wrong direction for the local dealers (70% of the sample), and is not statistically robust for the nonlocal dealers (the other 30%). Keep in mind that

this study *was* clearly well-designed to identify a possible failure of lab-field generalizability, but there is no statistically robust failure. Furthermore, the fact that the dealers offer *any* good-quality cards in response to price offers[17], as compared to offering the lowest quality for all offers, is evidence of some degree of reciprocity or other social preference.

Finally, this study is an important one because it has led to some broad, widely-publicized conclusions about human nature. A news article [Stanley, 2011] concluded that the 2006 card-dealer paper "helped refute a prevailing theory– which behavioral economists had arrived at in the lab– that humans were altruistic by nature". The article then quoted List as saying "Very few people will not screw each other. There are very few nice people out there."

(By the way, these conclusions about human altruism conflict sharply with the prevailing view of human exceptionalism in biology and anthropology. Scientists in those fields have much evidence that humans are more altruistic toward non-kin than all other species (except for a lively debate about prosociality of ape species). For example, Boyd and Richerson wrote: "Humans cooperate on a larger scale than most other mammals. Among social mammals, cooperation is mainly limited to relatives....The scale of human cooperation is an evolutionary puzzle" [Boyd and Richerson, 2009: 3281].[18]

Open air "flea" markets: List [2009] conducted a complex set of experiments about "open air" flea markets, which is something of a sequel to List [2006]. In these markets a vendor rents a venue and charges sellers a fee to set up a booth to sell both standardized and unstandardized goods, for a day or weekend. Buyers show up, and browse and haggle. List builds a staircase of institutional steps from the field setting, adding tighter experimental control at each step. The project exploits information from an unnamed "mole" informant seller who is colluding with other sellers to fix prices by an agreed-upon markup from the marginal cost paid to a common middleman.

In the lab and framed field experiments, buyer and seller values and costs are induced in

[17]Thanks to John Kagel for pointing this out.

[18]Boyd and Richerson's explanation for exceptional human altruism is that genes and culture coevolved in humans to create preferences for prosocial behavior and supporting adaptations (such as punishment).

the classic Smith-Plott design. In a natural field experiment, confederate buyer values are induced and the seller marginal costs are assumed (from the mole's inside information). Because the mole described clear price collusion, the data analysis is largely focused on whether there is cheating by undercutting the collusive price (as well as other related statistics such as price and allocative efficiency).

The closest lab-field designs are a FramedTable treatment in which worthless prop goods are traded for induced values, as compared to a Natural Field treatment in which actual items are traded (for the induced buyer value, and assumed seller marginal cost).

The key sample is 17 sellers who participated in either lab or field treatments, and also in the Natural Field treatments. While lab and field behavior are not always closely associated, List notes (without giving detail) that

> the best predictor of whether they will cheat in the natural field experiment is their measured cheating rate in the framed field treatments and the lab treatment with context [List, 2009: 44]

The overall deviation of transacted prices from collusive prices is 12.8% in FramedTable and 19% in Natural Field. Allocative efficiencies are 77% and 90%, respectively. Not enough details are reported to judge whether these modest differences are significant or not. One can therefore conclude that there is no reported evidence of significant differences in the lab and field in this paper, though there are small behavioral differences.

Donations to student funds: Benz and Meier [2008] compare naturally-occurring donations of CHF 7 and CHF 5 to two student fund, with experimental donations after classes from a CHF 12 endowment to the same two funds.

First note that even with this careful control, there are substantial differences in the lab and field treatments. The lab experiment was done at the end of a class, subjects were endowed with money to donate, and they could donate amounts in increments of .5 CHF. The field donations were done from their own money and were all or nothing.

The good news is that the average donation in the lab experiment was 9.46 CHF, compared to field donations of 9.00 and 9.50 in the four semesters before and after the lab experiment. The bad news is that the within-subject correlations between lab and field donation are modest, from .22 to .31.

In any case, while the lab-field correlation at the individual level is modest, the average donations in the two cases are very close. (In their 2007b paper, LL noted the low lab-field correlation across subjects, but did not mention the very close match of overall donations in the two conditions.) Is .30 a small or big lab-field correlation? Between-situation correlations of trait-like behavior are often not much larger than .30. There may be a practical upper limit on how much lab-field consistency we expect within people, *but the same limit applies to field-field comparisons* (see Liebbrandt [2011])

Soccer: Palacios-Huerta and Volij [2008] (PHV 2008 herein) designed a 2x2 lab matching pennies game designed closely to resemble soccer penalty kicks, in which players appear to approximately randomize [Palacios-Huerta, 2003]. In their simplified lab game a kicker and goalie choose left or right simultaneously. The game payoffs are kicker win rates estimated from 10 years of actual data from pro games.

They found that professional players approximate the mixed strategy equilibrium prediction in the lab games, with some modest deviations and serial correlation. College student subjects with no soccer experience deviated from equilibrium mixtures and had too many runs (a typical pattern). A sample of students who were playing in a serious amateur league played the lab games much like the professionals. This experiment shows a close match of play in the field and the lab by the professionals (though see Levitt, List, and Reiley [2009]; Wooders [2010]).

Communal fishing ponds: Stoop, Noussair and van Soest [2010] did a meticulous set of experiments on fishing and public goods contribution. Their baseline field setting is ponds in Holland where fishermen each pay a private fee, then have a communal pond stocked with

a specific number of extra fish that anyone can catch. They conduct a field experiment in which the incentives are changed. In their baseline condition the pond is stocked with 38 fish for 16 subjects, and they can only catch 2 fish per fisherman (in a 40-minute period). In a voluntary contribution (VCM) condition, the other fishermen in a 4-person group share a 6 euro bonus for each fish a specific subject catches under their 2-fish quota.

They record both catches and fishing effort (measured by the number of casts of their fishing lines per minute). The VCM condition has no effect on behavior. Comparable lab experiments are created with similar financial incentives (and contextual language). In those experiments there is more prosociality (under "fishing" to create earnings for others), and the fishermen are even more prosocial than students.

The apparent conclusion is that the economic prosociality observed in the lab is different than the selfish non-prosociality observed in the field, among fisherman. However, this ingenious study shows how difficult it is to carefully match all the features of lab and field settings. In the field an experimental period is 40 minutes. Fisherman make about three casts in five minutes, so it is easy for others to see very rapidly whether their peers are cooperating by casting less, so cooperation may dissolve quickly. In the lab a period is a couple of minutes and involves no real-time observation of what other people are doing. Underfishing in the field to create more euros for peers also means sitting idly by while others might fish (that is, there is a possible "boredom cost" which is not in induced in the lab equivalent).

Proofreading and exam grading: Boly [2011] did a beautiful simple study on responses of actual effort to incentives and monitoring in the lab and field. Subjects were instructed to find spelling errors in papers typed in French from dictated tapes. In the monitoring conditions either one or five out of 20 papers were examined and subjects were penalized if they miscounted the errors, based on the absolute deviation of detected and actual errors. In a high wage condition they were paid more (but no performance-linked pay). The lab subjects were students at CIRANO in Montreal. The field subjects were recruited in

Burkina Faso (mostly students at the local University of Ouagadougou) and told they were spell-checking "dicteés" (exams for civil service jobs). Note that the Ouagadougou subjects did not know they were in an experiment, but the Montreal lab subjects did.

Despite the large difference in subject background, the effects of monitoring and higher pay are generally identical up to two decimal places. The GLS coefficients of low and high monitoring on absolute deviations are -1.31 and -1.38 in the lab, and -1.33 and -1.49 in the field. There are also very close "period" effects (the order in which the monitored papers were done) on a slowdown in effort. And females are better graders in the lab (coefficient -.71) and the field (-.75). In a related report, Armantier and Boly [forthcoming] measure corruption rates by inserting a bribe into one of the exams. They find very close rates of corruption (acceptance of the bribe and grade misreporting) in the student and worker groups.

Fungibility of cash and in-kind subsidies: Abeler and Marklein [2010] did parallel field and lab experiments on fungibility of money. They compared overall spending and beverage spending when diners at a German wine restaurant were given identical-value 8 euro vouchers for either a specific good (beverages) or for an entire meal. Almost all diners spend at least 8 euros on beverages, so the restricted beverage voucher is essentially as valuable as the unrestricted voucher. They find that subjects spend more on the targeted goods in both the field restaurant setting, and in a stylized lab setting (with induced value consumption). The design does not permit a comparison of effect sizes, but is clear evidence that the sign and significance of the specific-voucher effect is similar in field and lab.

Summary: These studies on sports cards, markets, student donations, fishing, grading, and restaurant spending provide the closest matches of lab and field settings, protocols, and subjects available in the literature. Only one– fishing– shows poor generalizability of the essential behavior from the lab and field, which is statistically reliable. It therefore appears, empirically that when lab and field dimensions are carefully matched, good generalizability

is the rule and poor generalizability is the exception. A careful account of these studies therefore suggests a *general* conclusion about lab generalizability:

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).

This bold claim is meant to accomplish two goals: First, the default assumption in the economics profession should be that lab experiments *are* likely to generalize to closely matched field settings, until more clear studies show the opposite. This is the default assumption, and is generally supported by direct comparisons, in other fields such as biology studies comparing animal behavior in lab settings and in the wild.

Second, the claim is phrased boldly to attract the attention of people who think it is wrong. They should either bring existing examples to the attention of the profession, or do some diagnostic new studies which carefully match lab and field designs.

B. Apparent good generalizability with imperfect lab-field design match

In a few other studies, details of the lab design *were* specifically chosen to be highly comparable to a field setting. However, they are less conclusive than the studies in the previous section because either the subject pools are different or there are no field data on lab participants. In these studies, the lab-field match in effect size and direction is generally good, even though the lab environments have some of the features hypothesized to undermine generalizability.

Sports good and consumer good trading: List [2003] measured the strength of endowment effects in simple exchange experiments with sports paraphernalia traders (including professional dealers). The field data are artificial experiments conducted with participants who are endowed (randomly) with either good A or B and asked if they would trade it for the other good. Experienced traders and dealers trade a little less than half the time (45%), and inexperienced traders rarely trade (10-20%). There is a strong tendency for market

experience to reduce endowment effects.

However, there are substantial endowment effects in n-th-price auctions of sports card goods in field settings among dealers and nondealers. There appears to be an endowment effect even among dealers because the mean selling and buying prices are \$8.15 and \$6.27, a ratio of 1.30 (comparable to other ratios, e.g. Bateman et al. [1997]; though the ratio for nondealers is much, much higher at 5.6). The WTA-WTP difference is not significant (t=.87) but such a test does not have much power to find a typical empirical effect (with n=30 dealers in each group). Indeed, List [2003: footnote 11] says "I do not consider the point estimates herein to fully support neoclassical theory" (on the grounds that implied income elasticities for the endowed goods are "implausibly large".)

The paper also briefly reports a four-week experiment in which students come to a lab once a week, and each time are endowed with one of two different goods then allowed to trade for the other good (with different goods-pairs each week). In the first week only 12% trade away their endowed good, and 26% do in the fourth week. This trend is comparable to the experience effects that are associated with more trading in the field experiments.[19]

The strong experience effect in the lab experiment implies that the title of the paper, "Does market experience eliminate market anomalies?", is incomplete because it does not refer to the lab data at all. A more informative title would be, "Does *lab or* market experience eliminate anomalies?" The difference in language is important because the paper is typically cited as showing that field evidence from experienced market traders overturns an effect that is well-documented in simple lab settings. This rehearsed encoding of the result can easily lead to the mistaken impression that there are endowment effects in the lab but not among experienced traders in the field. That is certainly not the full conclusion since experience reduces endowment effects in *both* lab and field.

[19]Engelmann and Hollard [2010] essentially replicate this effect, by showing that forced trading reduces endowment effects (consistent with the interpretation that trading frequency mutes the effect of loss-aversion by shifting expectations of likely trade; see Della Vigna [2009: 328]. Feng and Seasholes [2005] and Dhar and Zhu [2006] report similar effects of investor sophistication and experience on reduction in the financial disposition effect (the tendency to oversell winning stocks and undersell losing stocks).

Semantics and titling can matter a lot since people probably remember the title or the gist of the paper better than its detail (if they even read it all), since gist-retention is how unaided memory typically works. In fact, none of the more widely-cited papers[20] which are reported as citing List [2003] mention the student lab data *at all*. For example, Haigh and List [2005: 524] write:

> "in light of some recent studies (e.g., List [2002; 2003; 2004]) that report market anomalies in the realm of riskless decision-making are attenuated among real economic players who have intense market experience, the current lot of experimental studies and their support of MLA [myopic loss-aversion] may be viewed with caution."

Note that their passage distinguishes between "real economic players" and "experimental studies", as if the results are fundamentally different in those groups. But the List [2003] paper actually shows behavior that is fundamentally *the same* in those two groups.

Event partitions in prediction markets: Sonneman et al. [2011] studied whether a simple bias in probability judgment is manifested in lab experiments, field experiments and field data. The bias is that when a continuous variable is partitioned into N sets, judged probabilities over sets tend to be biased toward $1/N$. For example, if the outcomes of the NBA playoff finals, measured by total games won, are partitioned into (0,3), (4-6), (7+) the combined judged probability of the first two sets is larger than the judged probability of the "packed" set (0-6) when the partition is (0-6), (7+). A common intuition is that self-selection into field markets, stakes, and experience could reduce or erase these types of psychological effects.

To test this hypothesis Sonneman et al. compared two-hour lab markets for naturally-occurring events, seven-week field experiments on NBA and World Cup outcomes, field data

[20]The most widely-cited are defined for this purpose as those with more than 100 Google Scholar cites, as of 24 April 2011)

from 150 prediction markets for economics statistics, and data on betting odds for a million horse races [Snowberg and Wolfers, 2010].

Some partition-dependence is evident in all four markets. Statistical analysis indicates that the probabilities implied by economic statistics market represent a mixture of a $1/N$ prior belief with judged likelihood, much as in lab experiments. The combination of trader self-selection into those markets, its special context, and incentive does not undermine the generalizability of the lab result.

Sharing with charity: An experiment on dictator game sharing with a specific Chinese poverty charity showed a similar drop when the income endowment was earned vs. unearned in a student lab experiment vs. citizen customers outside a supermarket [Carlsson, He, Martinsson, 2010]. The students did give more in general, but that effect is confounded with lab-field, demographic differences, and some procedural variables. This is an illustration of Kessler and Vesterlund's (this volume) important point that the lab and field could generate different behavioral level effects, but also generate comparable comparative static responses to design variables.

Swedish lotteries: Östling et al. [2011] collected data from a daily Swedish lottery in which players choose integers from 1 to 99,999, and the lowest unique integer wins a large prize. This *l*owest *u*nique *p*ositive *i*nteger (LUPI) game provides a rare sharp test of mixed strategy equilibrium in the field. The Poisson-Nash equilibrium predicts comparable numbers of choices of numbers 1-5000 with a sharp dropoff after that point. Compared to that benchmark, there are too many low numbers, not enough numbers in the range (3000-5000) and too many higher numbers. The data are fit reasonably well by a QR cognitive hierarchy model with average thinking $\tau = 1.80$ (close to a typical estimate from many lab experiments; e.g., Camerer, Ho, and Chong [2004]).

Östling et al. then designed a lab experiment to replicate the key theoretical features as closely as possible, while scaling down the number of players and the integer range. The

central empirical feature of the field data-too many low and high numbers– also occurs in the lab data. This is an example of when qualitative results are a close match but quantitative details are not.

Silent auctions: Isaac and Schnier [2006] compare field data from three "silent auctions" with lab experiments designed to be closely comparable. They focus on the frequency of "jump bids" which top previous bids by more than the minimum increment. The frequency of jump bids is 9-39% in the field and 40-61% in the lab. Jumping one's own bid is rare in the field (0%) and in the lab (.09-.16%). Thus, the magnitudes of these two simple statistics are suggestive of modest comparability. They also run probit regressions of what variables predict jump bids and find substantial overlap in sign and significance (their Table 9). In only one case is there a significant effect in the field-less jump bidding early in the auctions-which is significant and opposite in sign in the lab.

"Deal or No Deal": Post et al. [2008] started with field data from the "Deal or No Deal" television show in three countries (US, Germany, Holland). After making many simplifying assumptions, they could infer risky choice preference parameters from the contestants' decisions. The central interesting feature of their analysis is that decisions are consistent with risk tastes depending on prior outcomes in a sequence of choices. Specifically, contestants exhibit a "break-even effect" (taking risks to reach a previous point of reference) and a "house money" effect (taking more risks when the set of possible winning amounts shifts upward). They also conducted parallel lab experiments which had many features of the TV game show, including a grinning "game show host" (a popular lecturer at Erasmus University), live audience, video cameras, and a computerized display of unopened briefcases, remaining prizes and "bank offers". They report that "the audience was very excited and enthusiastic during the experiment, applauding and shouting hints, and most contestants showed clear symptoms of distress".

Post et al. concluded that

> choices in the experiment are remarkably similar to those in the original TV show, despite the fact that the experimental stakes are only a small fraction of the original stakes. Consistent with the TV version, the break-even effect and the house-money effect also emerge in the experiments. [Post et al., 2008: 68]

The Weakest Link: In the game show "The Weakest Link" people take turns answering trivia questions, and participants can exclude others based on previous performance. Antonovics, Arcidiacono, and Walsh [2009] compared field data from the TV show with a similar lab experiment, to see whether opponent gender makes a difference. They find that in general, men do better against female opponents. However, the effect disappears in younger contestants. The effect is also absent in younger college students in a lab environment with higher stakes. Thus, going from a particular field setting to a comparable lab setting yields the same results, provided age and motivation are controlled for.

C. Lab-field generalization with substantial design and subject differences

A bolder, and more common, type of lab-field generalization correlates individual lab-based measurements of behavior or preferences with naturally-occurring measures of socioeconomic activity from the same or different individuals (or sometimes aggregated at the professional, community, firm or national level). These analyses are not guaranteed to produce correlation, even if there is no fundamental lab-field difference ceteris paribus, because there are differences in the lab and field settings.

For example, Barr and Serneels [2004] did trust experiments with Ghanian manufacturing workers, whose characteristics, output and earnings were measured in a typical survey. In the experiments, workers generally repaid 1, 1.5, or 2 times as much as was invested by first-movers. Define those repaying more than 1.5 as "high reciprocators". They find that across firms, the proportion of high reciprocators is strongly correlated (r=.55, p $<$ .01) with output per worker. High-reciprocator workers also earn 40% more (t=2.7). However, instrumental control for reverse causality from earnings to reciprocity weakens the effect to an insignificant 17% increase (t=.28).

The results from recent studies of this type typically find positive associations between lab and field measures. I briefly mention findings from a few studies organized by domain. My characterization of the studies' findings is meant to be not too aggressive about positive results (many are statistically significant, though not large in magnitude). More formal analyses should therefore certainly be conducted.

Pricing:

- Comparable price elasticities (-.69 and -.79) in door-to-door field and lab sales of strawberries [Brookshire, Coursey, and Schulze, 1987]

- Actual market share of "antibiotic-friendly" pork was not significantly different than a forecast based on a choice experiment outside a store in Oklahoma [Lusk, Pruitt, and Norwood, 2006].

Risk and time preference:

- Impatience and risk-tolerance are correlated with contribution to pensions by the self-employed in Chile [Barr and Packard, 2000].

- On a TV game show a simple game involving one or two spins of a wheel. In the field data and in a lab replica, subjects fail to take a second spin frequently when they should spin, in both field and lab [Tenorio and Cason, 2002].

- Bidders in a lab "Price is Right" replica exhibit two anomalies which are also observed in field data from the TV game show [Healy and Noussair, 2003]: The last bidder does not optimize 36% of the time in the lab, vs. 43% in the field; and bids decline (as predicted by theory) only 7.7% of the time in the lab vs. 12.1% in the field.

- Borrowers who are present-biased experimentally have larger credit card balances [Meier and Sprenger, 2008].

- Individual measures of time discounting from experimental amount-delay choices correlate modestly but reliably with various field data on self control (diet, exercise, saving, gambling, etc.) [Chabris et al., 2008]. Correlations improve with aggregation.

- Female Connecticut trick-or-treaters who make an ambiguity-averse choice of a candy bag are more likely to wear a "less risky" (more popular) Halloween costume [Anagol et al., 2010]

- In price-list tests of risk-aversion and time discounting in artificial lab experiments, using both student groups and citizen volunteers from a Danish representative sample, Andersen, Harrison, Lau, and Rutstrom [2010] "find no significant difference in the average degree of risk aversion and discount rates between the field [citizen] and laboratory samples." They also note that there is a common experimenter effect on discount rates in the two samples, and suggest it might be a "priming" effect in which a slower experimenter induced more patient choices.

- Slow adoption of modified Bt cotton by Chinese farmers is correlated with risk- and loss-aversion, and lower overweighting of low probability [Liu, 2011]

Peer effects:

- A numerical estimate of the strength of peer effects on output in laboratory envelope-stuffing is "very similar to a comparable estimate derived by Ichino and Maggi [2000] with observational data" [Falk and Ichino, 2006]

Prosociality:

- Peruvian villagers who were less trustworthy players in trust experiments also defaulted on microfinance loans at higher rates [Karlan, 2005]. However, the link between trust and loan repayment is less clear because trust can reflect altruistic giving, or an expectation of repayment. Karlan concludes [p. 1698]. "This endorses experimental

economics as a valid measurement tool for field research, and the Trust Game as a valid method to measure trustworthiness, but **not** as a method to measure trust".

- Conditional cooperation and verbal disapproval in a public goods game predicts group fishing productivity in Japan [Carpenter and Seki, 2005].

- Giving in an experimental dictator game with charity recipients predicts whether Vermonters volunteer to fight fires [Carpenter and Myers, 2007]

- Experimental public goods contributions and patience for water predict limited (prosocial) common pool resource extraction among Brazilians who catch fish and shrimp [Fehr and Liebbrandt, 2008]

- An experimental bribery game done with Oxford students from around the world (in both an original sample and a replication) found that experimental bribery and acceptance among undergraduates, but not graduate students, was correlated with an index of corruption (Transparency International) in a students home country [Barr and Serra, 2010].

- Experimental trustworthiness among University of Chicago MBAs correlates with their individual donations to a class gift [Baran, Sapienza, and Zingales, 2010].

- Dictator game allocations from Ugandan teachers to parents correlate with the teachers' actual teaching time (the inverse of absenteeism) [Barr and Zeitlin, 2010]

- Effort in a lab gift-exchange experiment with students in Lyon, France is positively correlated with both worker income and the worker's income rank in a comparison group [Clark, Masclet, and Villeval, 2010]. The same correlation signs and statistical strengths were also observed in ISSP survey questions from 17 OECD countries about willingness to "work harder than I have to in order to help the firm or organization I work for to succeed", as correlated with reported income and income rank.

- Relative favoritism of a low-status outgroup (the Khmer) by Vietnamese and Chinese in sharing and cooperation decisions is consistent with government policies favoring the Khmer (e.g., education and tax subsidies) [Tanaka and Camerer, 2010] (no data are available at the individual level on policy preferences, however).

- Left-handedness among women is correlated with selfishness in a lab dictator game and with measures of charitable donation in field surveys [Buser, 2010]. Similarly, men are more trusting and reciprocal in a lab trust game and endorse the same behaviors in a LISS survey. (Some other lab behaviors and field measures do not show much association, however.)

- Group-level conditional cooperation in experiments is correlated with success in managing forest commons in Ethiopia [Rustagi, Engel, and Kosfeld, 2010].

- Experimental public goods (PG) contributions in 16 Indian villages correlate with how much salt each villager took from a common pool when he or she arrived to collect experimental earnings [Lamba and Mace, in press]. The individual-level correlations between PG contribution and salt-taking are low (r=.057) but the village-level correlations are extremely high (r=.871).[21] This study is an important empirical reminder that the level at which lab-field generalizability is best is not necessarily the individual level.

- Laboratory measures of prosocial behavior among truckers is correlated field prosocial behavior under comparable conditions (anonymous unrepeated interactions) [Anderson et al., 2011].

- Choices by Ethiopian nurses and doctors to work for NGOs correlated with their experimental prosociality in a generalized trust game (where a Responder can share money

[21]Their paper also has excellent data on individual-level characteristics which is important for drawing conclusions between ethnic groups and villages. See also Henrich et al. [in press] for a comment disputing some of Lamba and Mace's conclusions.

that one Proposer invested with a different Proposer). The correlation exists even after controlling for self-reported desire to help the poor [Serra, Serneels and Barr, 2011].

- A low-income Dallas population who contribute more in public goods experiments are more prosocial in other experiments, and self-report more donation and volunteering outside the lab [de Oliveira, Croson, and Eckel, 2011].

- Students making choices online about donating mosquito nets to pregnant mothers in Kenya from lottery winnings, and donating to another anonymous student, make choices which are correlated across the two types of donees [Coffman, forthcoming].

Two studies show lab comparisons that go in the opposite direction of a comparable field comparison:

- In theory and in lab experiments, all-pay auctions yield more revenue than public or anonymous voluntary contribution mechanisms (e.g., Schram and Onderstal [2009]). However, the opposite pattern was found (all-pay auctions yield the least) in a large door-to-door fundraising experiment with 4500 households [Onderstal, Schram, and Soetevent, 2011]. Note, even further, that in their field experiment Onderstal et al. [2011] found that VCM raised more revenue than lotteries, the opposite result of a field experiment by Landry et al. [2006]. This observation is a simple reminder that field-field generalizability does not always work either: While the Onderstal et al. lab and field results do not match up well, neither do the results of the two different field experiments.[22]

- Results of corruption experiments conducted in Australia, India, Indonesia, Singapore do not strongly correlate with national corruption indices [Cameron et al., 2009]. Lab and field corruption are both high in India and low in Australia, but the lab and field results are mismatched in Indonesia and Singapore. The authors suggest that recent *changes* in public attitudes toward corruption could account for the two mismatches.

[22]Thanks to Arthur Schram for pointing this out.

Several papers compare general regularities derived from field data (typically about the response to institutional changes) with highly stylized lab equivalents. In these cases there is a clear hope for generalizability in conducting the lab experiments, but the lab-field subject task and identity are not closely matched. However, the general finding is that comparative statics responses to institutional changes, and general regularities, often are similar in sign and magnitude in the field.

Kagel and Levin [1986] show correspondence to many properties of lab common-value auctions, showing overbidding and a "winner's curse", and patterns in drainage leases in the Gulf of Mexico. Kagel and Roth [2000] created experimental matching markets with different features that are motivated by differences in observed behavior in naturally-occurring markets, and find parallel behavior in the lab and field. Blecherman and Camerer [1996] observed parallels between overbidding for baseball free agents (using field data) and overbidding in highly stylized lab experiments with features of free agency bidding. Bolton, Greiner, and Ockenfels [2009] study changes in online auction reputation systems in simplified lab settings and show that they are similar to field responses.

6 Conclusion: Where do we go from here?

This paper considers the issue of whether economic lab experiments should be expected to generalize to specific naturally-occurring field settings. I suggest that generalizability of lab results is an exaggerated concern among non-experimenters for three possible reasons.

First, the scientific perspective that governed experimental economics from the beginning [Smith, 1976, 1982] is that all empirical methods are trying to accumulate regularity about how behavior is *generally* influenced by individual characteristics, incentives, endowments, rules, norms, and other factors. A typical experiment therefore has no specific target for "external validity"; the "target" is the general theory linking economic factors to behavior. (That's also the same "target" a typical field study has.) A special concern for external valid-

ity is certainly appropriate when the only goal of an experiment is to provide guidance about how behavior might work in a specific external setting– in Julian Stanley's "known circumstances" education language. But such targeted guidance is rarely the goal of experiments in economics.

Second, when experiments are criticized for limited generalizability (as by LL in many passages), that criticism depends on contrasting stereotypes of a canonical low-stakes, artificial experiment with students and a canonical field setting with self-selected skilled agents and high stakes. Criticisms that depend on these contrasted stereotypes ignore the crucial fact that experiments can be very different and that *more experiments can always be conducted*. Since many different types of experiments can be run, the threats to external validity that LL note– moral considerations, obtrusiveness, context, self-selection, and stakes– can typically be varied in experiments to see how much they matter (as Falk and Heckman [2009] noted too).

Third, non-experimenters in economics often do not realize the extent to which modern economics experiments, which developed in the 1970s, differ very sharply from methods in stereotypical psychology experiments from the 1960s and beyond. The Schultz [1969] concept, quoted by LL, that the experimenter-subject relationship is "matched only by that of...drill sergeant and trainee" is simply not an accurate description of modern experimental economics experiments. It is certainly true that in a stereotypical *psychology* experiment, there is often no clear performance metric and no incentive pay (students often must participate for "course credit"), there is sometimes deception, and fears about demand effects are reasonable since the experimenters often want the experiment to "work" (i.e. give a particular outcome). But these features are *not* typical of economics experiments.

Let's end by recalling the ominous warning in Levitt and List's extensive writing about lab-generalizability:

> Yet unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely

> converge across the lab and many field settings. [LL, 2007a: 364]

Fortunately, their scary warning seems to be wrong about the most direct lab-field comparisons and is certainly not true in most indirect lab-field comparisons, including those generalizing prosociality from the lab to the field.

Tables and Figures

Table 1: Coefficient from Tobit regression of quality on price and estimated gift exchange coefficient

	Lab-context	Lab-market	Field	Field (announce grading)	Field (grading)	Field (pooled)
	Described as cards	cards	cards	tickets	tickets	tickets
Price coefficient	.05 (4.3)	.02 (4.4)	.02 (6.6)	.02 (2.1)	.02 (1.1)	.02 (2.6)
Gift exchange estimate θ	$.65 (4.7)	$.45 (2.1)	$.21 (5.0)	$.17 (1.1)	$.23 (1.1)	$.19 (2.3)
Buyers incentivized?	Yes	No	No	No	No	No
N	32	60	100	54	36	90

Source: List [2006]

Note: T-statistics in parentheses. Dealer random effects included. The column headings above correspond to original [List, 2006: Table 4] headings of: Lab-context, Lab-market, Floor (cards), Floor-Announce grading, Floor-Grading.

Table 2: Mean quality supplied in response to higher prices (reciprocity) in adjacent lab and field markets by dealer sample

	Non-local		Local	
	Mean (std. dev.)	N	Mean (std. dev.)	N
Lab	1.11 (1.17)	9	.95 (1.32)	21
Field	.13 (.92)	15	1.40 (.81)	35
Nonparametric tests: statistic, two-tailed p-value				
Mann-Whitney	38, .084		461.5, .114	
Fligner-Policello	2.09, .020		1.45, .073	
Epps-Singleton	3.83, .429		22.07, .0002	

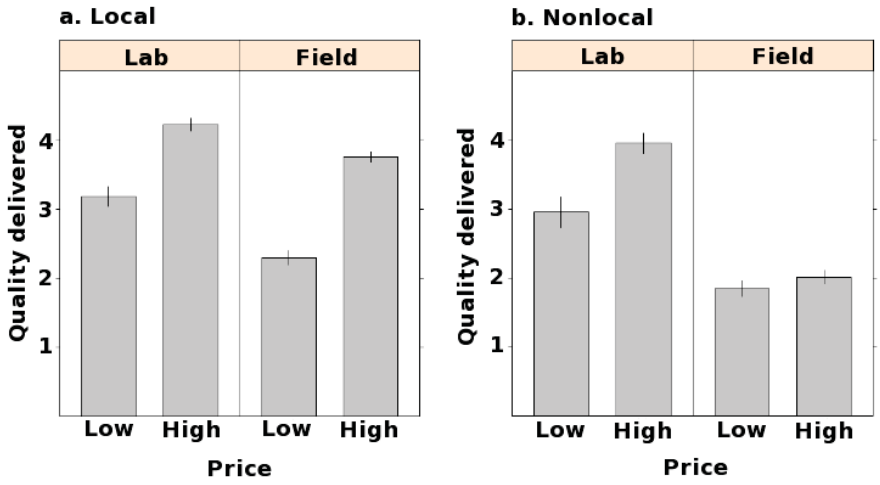

Figure 1: Quality responses to low and high price offers. **a.** Local dealers (lab, left; field, right). **b.** Nonlocal dealers (lab, left; field, right). Adapted from: List [2006].

References

Abeler, J. and F. Marklein. 2010. Fungibility, Labels and Consumption. *IZA Discussion Paper* No. 3500

Alatas, V., L. Cameron, A. Chaudhuri, N. Erkal, and L. Gangadharan. 2009. Subject Pool Effects in a Corruption Experiment: A Comparison of Indonesian Public Servants and Indonesian Students. *Experimental Economics* 12(1): 113-132.

Anagol, S., S. Bennett, G. Bryan, T. Davenport, N. Hite, D. Karlan, P. Lagunes, and M. McConnell. 2010. There's Something about Ambiguity. *Working paper.*

Andersen, S., S. Ertac, U. Gneezy, M. Hoffman, and J. A. List. In press. Stakes Matter in Ultimatum Games *American Economic Review*

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2010. Preference Heterogeneity in Experiments: Comparing the Field and Laboratory. *Journal of Economic Behavior & Organization* 73(2): 209-24.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutstrm. 2008. Risk Aversion in Game Shows. In *Risk Aversion in Experiments: Research in Experimental Economics, Vol. 12* eds. G. W. Harrison and J. Cox. 359-404. Emerald Group Publishing/JAI Press.

Anderson, J., M. Bombyk. S. Burks, J. Carpenter. D. Ganzhorn. L. Goette, D. Nosenzo, and A. Rustichini. 2011. Lab Measures of Other-Regarding Behavior Predict Some Choices in a Natural On-the-Job Social Dilemma: Evidence from Truckers. *Working paper.*

Anderson, J., S. Burks, J. Carpenter, L. Goette, K. Maurer, D. Nosenzo, R. Potter, K. Rocha, and A. Rustichini. 2010. Self Selection Does Not Increase Other Regarding Preferences among Adult Laboratory Subjects, but Student Subjects May Be More Self-regarding than Adults. *Working paper.*

Andreoni, J. and B. Bernheim. 2009. Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica* 77(5): 1607-1636.

Angrist, J. and A. Krueger. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*

15(4): 6985.

Antonovics, K., P. Arcidiacono, and R. Walsh. 2009. The Effects of Gender Interactions in the Lab and in the Field. *Review of Economics and Statistics* 91(1): 152-162.

Armantier, O. and A. Boly. Forthcoming. A Controlled Field Experiment on Corruption. *European Economic Review*

Ball, S. B. and P.-A. Cech. 1996. Subject Pool Choice and Treatment Effects in Economic Laboratory Research. In *Research in Experimental Economics, Vol. 6* ed. R. Mark Isaac. 239-292. Elsevier Science and Technology Books.

Banerjee, A. and E. Duflo. 2009. The Experimental Approach to Development Economics. *Annual Review of Economics* 1: 151-178.

Baran, N. M., P. Sapienza, and L. Zingales. 2010. Can We Infer Social Preferences from the Lab? Evidence from the Trust Game. *NBER Working Papers*.

Bardsley, N. 2005. Experimental Economics and the Artificiality of Alteration. *Journal of Economic Methodology* 12: 239-251.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2009. *Experimental Economics: Rethinking the Rules* Princeton University Press.

Barr, A. and T. Packard. 2000. Revealed and Concealed Preferences in the Chilean Pension System: An Experimental Investigation. *Department of Economics Discussion Paper Series*. University of Oxford.

Barr, A. and P. Serneels. 2004. Wages and Reciprocity in the Workplace. *Centre for the Study of African Economies Series*.

Barr, A. and D. Serra. 2010. Corruption and Culture: An Experimental Analysis. *The Journal of Public Economics* 94(11-12): 862-869.

Barr, A. and A. Zeitlin. 2010. Dictator Games in the Lab and in Nature: External Validity Tested and Investigated in Ugandan Primary Schools. *CSAE Working Paper Series*. Centre for the Study of African Economies, University of Oxford.

Bateman, I., A. Munro, B. Rhodes, C. Starmer, and R. Sugden. 1997. A Test of the

Theory of Reference-Dependent Preferences. *Quarterly Journal of Economics* 112(2): 479-505.

Belot, M., R. Duch, and L. Miller. 2010. Who Should be Called to the Lab? A Comprehensive Comparison of Students and Non-students in Classic Experimental Games. *Nuffield centre for experimental social sciences, University of Oxford, Discussion paper series.*

Benz, M. and S. Meier. 2008. Do People Behave in Experiments as in the Field?– Evidence from Donations. *Experimental Economics* 11(3): 268-281.

Blecherman, B. and C. F. Camerer. 1996. Is There a Winner's Curse in the Market for Baseball Players? Evidence from the Field. *Social Science Working Paper*. California Institute of Technology.

Bolton, G., B. Greiner, and A. Ockenfels. 2009. Engineering Trust– Reciprocity in the Production of Reputation Information. *Working Paper Series in Economics*. University of Cologne.

Bolton, G. E. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* 90(1): 166-193.

Boly, A. 2011. On the Incentive Effects of Monitoring: Evidence from the Lab and the Field. *Experimental Economics* 14(2): 241-253.

Boyd, R. and P. J. Richerson. 2009. Culture and the Evolution of Human Cooperation. *Philosophical Transactions of the Royal Society B* 364: 3281-3288.

Bracht, G. H. and G. V. Glass. 1968. The External Validity of Experiments. *American Educational Research Journal* 5: 437-474.

Brookshire, D. S, D. L. Coursey, and W. D. Schulze. 1987. The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior. *Economic Inquiry* 25(2): 239-250.

Buser, T. 2010. Handedness Predicts Social Preferences: Evidence Connecting the Lab to the Field. *Tinbergen Institute Paper* TI 2010-119/3.

Camerer, C. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*

Princeton: Princeton University Press.

Camerer, C., T-H. Ho, and J-K. Chong. 2004. A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics* 119(3): 861-898.

Camerer, C. and R. Hogarth. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty* 19(1-3): 7-42.

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

Cameron, L., A. Chaudhuri, N. Erkal, and L. Gangadharan. 2009. Propensities to Engage in and Punish Corrupt Behavior: Experimental Evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics* 93(7-8): 843-51.

Campbell, D. and J. Stanley. 1963. Experimental and Quasi-Experimental Designs for Research on Teaching. In *Handbook of Research on Teaching* ed. N. L. Gage. Chicago: Rand McNally.

Carlsson, F., H. He, and P. Martinsson. 2010. Windfall vs. Earned Money in the Laboratory: Do They Affect the Behavior of Men and Women Differently? *Working Papers in Economics*. University of Gothenburg.

Carpenter, J. and C. K. Myers. 2007. Why Volunteer? Evidence on the Role of Altruism, Reputation, and Incentives. IZA Discussion Papers.

Carpenter, J. and E. Seki. 2005. Competitive Work Environments and Social Preferences: Field Experimental Evidence from a Japanese Fishing Community. *Middlebury College Working Paper Series*. Middlebury College.

Chabris, C. F., D. I. Laibson, C. L. Morris, J. P. Schuldt, and D. Taubinsky. 2008. Individual Laboratory-measured Discount Rates Predict Field Behavior. *Journal of Risk and Uncertainty* 37: 237-269.

Clark, A., D. Masclet, M. C. Villeval. 2010. Effort and Comparison Income. Experimental and Survey Evidence. *Industrial and Labor Relations Review* 63(3): 407-426.

Cleave, B., N. Nikiforakis, and R. Slonim. 2011. Is There Selection Bias in Laboratory Experiments? The Case of Social and Risk Preferences. *IZA Discussion Paper* No. 5488.

Clement, D. 2002. Interview with Gary Becker. In *The Region* The Federal Reserve Bank of Minneapolis.

Coffman, L. C. Forthcoming. Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics*

Deaton, A. 2009. Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. In *NBER Working Paper* National Bureau of Economic Research.

Della Vigna, S. 2009. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature* 47: 315-372.

de Oliveira, A., R. Croson, and C. C. Eckel. 2011. The Giving Type: Identifying Donors. *Journal of Public Economics* 95(5-6): 428-435.

Dhar, R. and N. Zhu. 2006. Up Close and Personal: Investor Sophistication and the Disposition Effect. *Management Science* 52(5): 726-740.

Eckel, C. C. and P. J. Grossman. 2000. Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method in Dictator Experiments. *Experimental Economics* 3(2): 107-120.

Engelmann, D. and G. Hollard. 2010. Reconsidering the Effect of Market Experience on the "Endowment Effect. *Econometrica* 78(6): 2005-2019.

Ensminger, J. E. 2004. Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* eds. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. London: Oxford University press.

Falk, A. and J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* 326(5952): 535-538.

Falk, A. and A. Ichino. 2006. Clean Evidence on Peer Effects. *Journal of Labor Eco-*

nomics 24(1): 39-57.

Fehr, E. and A. Leibbrandt. 2008. Cooperativeness and Impatience in the Tragedy of the Commons. *IZA Discussion Papers*. 3625.

Feng, L. and M. Seasholes. 2005. Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets? *Review of Finance* 9(3): 305-351.

Frchette, G. R. Forthcoming. Laboratory Experiments: Professionals versus Students. In *The Methods of Modern Experimental Economics,* eds. Guillaume Frchette and Andrew Schotter. Oxford University Press.

Haigh, M. and J. A. List. 2005. Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis. *Journal of Finance* 60(1): 523-534.

Healy, P. J. and C. Noussair. 2004. Bidding Behavior in the Price is Right Game: an Experimental Study. *Journal of Economic Behavior & Organization* 54(2): 231-247.

Henrich, J., R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, et al. 2005. "Economic Man" in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies. *Behavioral and Brain Sciences* 28(6): 795-815.

Henrich, J., R. Boyd, R. McElreath, M. Gurven, P. J. Richerson, J. Ensminger, M. Alvard, A. Barr, C. Barrett, A. Bolyanatz, C. Camerer, J. C. Cardenas, E. Fehr, H. Gintis, F. GilWhite, E. Gwako, N. Henrich, K. Hill, C. Lesorogol, J. Q. Patton, F. Marlowe, D. Tracer and J. Ziker. In press. Culture Does Account for Variation in Game Behaviour. *Proceedings of the National Academy of Sciences*.

Henrich, J., R. McElreath, A. Barr, J. Ensimger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, J. Ziker. 2006. Costly Punishment Across Human Societies. *Science* 312: 1767-1770.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences* 33: 61-83.

Hoffman, E., K. A. McCabe, and V. L. Smith. 1998. Behavioral Foundations of Reci-

procity: Experimental Economics and Evolutionary Psychology. *Economic Inquiry* 36(3): 335-352.

Hoffman, M. H. and J. Morgan. 2011. Who's Naughty? Who's Nice? Social Preferences in Online Industries. *Working paper*.

Ichino, A. and G. Maggi. 2000. Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. *Quarterly Journal of Economics* 115(3): 1057-1090.

Imbens, G. 2009. Better Late Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *NBER Working Papers Series* National Bureau of Economic Research.

Isaac, R. M. K. and K. Schnier. 2006. Sealed Bid Variations on the Silent Auction. In *Experiments Investigating Fundraising and Charitable Contributors (Research in Experimental Economics, Vol. 11)* eds. R. M. Isaac and D. D. Davis, 31 - 46. Emerald Group Publishing Limited.

Kagel, J. and D. Levin. 1986. The Winner's Curse and Public Information in Common Value Auctions. *American Economic Review* 76: 894-920.

Kagel, J. H. and A. E. Roth. 2000. The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment. *Quarterly Journal of Economics* 115(1): 201-235.

Karlan, D. S. 2005. Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review* 95(5): 1688-1699.

Krupka, E. and R. Weber. 2008. Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *IZA Discussion Paper* Institute for the Study of Labor. No. 3860.

Lamba, S. and R. Mace. In press. Demography and Ecology Drive Variation in Cooperation across Human Populations. *Proceedings of the National Academy of Sciences*.

Lambdin, C. G. and V. A. Shaffer. 2009. Are within-Subjects Designs Transparent?

Judgment and Decision Making 4(7): 554-566.

Landry, C., A. Lange, J. List, M. Price, and N. Rupp. 2006. Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment. *Quarterly Journal of Economics* 121(2): 747-782.

Lazear, E. P., U. Malmendier, and R. A. Weber. Forthcoming. Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics*

Levitt, S. and S. Dubner. 2009. *Super Freakonomics* HarperCollins.

Levitt, S. and J. A. List. 2007a. Viewpoint: On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics* 40(2): 347-370.

Levitt, S. and J. A. List. 2007b. What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World. *Journal of Economic Perspectives* 21(2): 153-174.

Levitt, S. and J. A. List. 2008. Homo Economicus Evolves. *Science* 319(5865): 909-910.

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

Liebbrandt, A. 2011. The External Validity of the Public Goods, Trust, Dictator, and Ultimatum Game. Presented at the Annual International Meeting of the Economic Science Association. Chicago, IL.

List, J. A. 2002. Preference Reversals of a Different Kind: The More is Less Phenomenon. *American Economic Review* 92(5): 1636-1643.

List, J. A. 2003. Does Market Experience Eliminate Market Anomalies? *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. 2006. 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. 2007. On the Interpretation of Giving in Dictator Games. *Journal of Political*

Economy 115(3): 482-494.

List, J. A. 2009. The Economics of Open Air Markets. *NBER Working Paper*.

Liu, E. M. 2011. Time to Change What to Sow: Risk Preferences and Technology Adoption Decisions of Cotton Farmers in China. *Working paper*.

Lusk, J. L., F. B. Norwood, and J. R. Pruitt. 2006. Consumer Demand for a Ban on Antibiotic Drug Use in Pork Production. *American Journal of Agricultural Economics* 88(4): 1015-1033.

Meier, S. and C. Sprenger. 2010. Present-Biased Preferences and Credit Card Borrowing. *American Economic Journal: Applied Economics* 2(1): 193-210.

Onderstal, S., A. Schram, and A.d Soetevent. 2011. Bidding to Give in the Field: Door-to-Door Fundraisers Had It Right from the Start. *Unpublished work*.

Oosterbeek, H., R. Sloof, and G. van de Kuilen. 2004. Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics* 7(2): 171-188.

Orne, M. T. 1962. On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *American Psychologist* 17(11): 776-783.

Östling, R., J. T. Wang, E. Y. Chou, and C. F. Camerer. 2011. Testing Game Theory in the Field: Swedish LUPI Lottery Games. *American Economic Journal: Microeconomics* 3(3): 1-33.

Palacios-Huerta, I. 2003. Professionals Play Minimax. *Review of Economic Studies* 70(2): 395-415.

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

Pierce, A. H. 1908. The Subconscious Again. *Journal of Philosophy, Psychology and Scientific Methods* 5(10): 264-271

Post, T., M. J. van den Assem, G. Baltussen, and R. H. Thaler. 2008. Deal or No Deal?

Decision Making under Risk in a Large-Payoff Game Show. *American Economic Review*, 98(1): 38-71.

Pruitt, D. and M. Kimmel. 1977. Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future. *Annual Revue of Psychology* 28: 363-392.

Rabin, M. 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83(5): 1281-1302.

Rapoport, A. 1970. Conflict Resolution in the Light of Game Theory and Beyond. In *The Structure of Conflict*, ed. P. Swingle. New York: Academic Press.

Rebitzer, J. B. 1988. Unemployment, Labor Relations and Unit Labor Costs. *American Economic Review* 78(2): 389-394.

Rustagi, D., S. Engel, and M. Kosfeld. 2010. Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management. *Science* 330(6006): 961-965.

Samuelson, P. and W. Nordhaus. 1985. *Economics* New York: McGraw-Hill.

Schram, A. and S. Onderstal. 2009. Bidding to Give: An Experimental Comparison of Auctions for Charity. *International Economic Review* 50(2): 431-457.

Schultz, D. P. 1969. The Human Subject in Psychological Research. *Psychological Bulletin* 72(3): 214-228.

Serra, D., P. Serneels, and A. Barr. 2011. Intrinsic Motivations and the Non-profit Health Sector: Evidence from Ethiopia. *Personality and Individual Differences* 51(3): 309-314.

Sims, C. 2010. But Economics Is Not an Experimental Science. *Journal of Economic Perspectives* 24(2): 59-68.

Smith, V. 1982. Microeconomic Systems as an Experimental Science. *American Economic Review* 72(5): 923-955.

Smith, V. and J. W. Walker. 1993. Monetary Rewards and Decision Cost in Experimental Economics. *Economic Inquiry* 31(2): 245-261.

Snowberg, E. C. and J. Wolfers. 2010. Explaining the Favorite-Longshot Bias: Is It Risk-Love or Misperceptions? *CEPR Discussion Papers*.

Sonnemann, U., C. F. Camerer, C. R. Fox, and T. Langer. 2011. Partition Dependence in Prediction Markets: Field and Lab Evidence. *Working paper*. University of Muenster.

Stanley, O. 2011, February 22. Chicago Economist's 'Crazy Idea' Wins Ken Griffins Backing. *Bloomberg Markets Magazine*. http://www.bloomberg.com/news/2011-02-23/chicago-economist-s-crazy-idea-for-education-wins-ken-griffin-s-backing.html.

Stoop, J., C. Noussair, and D. van Soest. 2010. From the Lab to the Field: Public Good Provision with Fishermen. *Manuscript in progress*.

Tanaka, T. and C. F. Camerer. 2010. Patronizing Economic Preferences toward Low-Status Groups in Vietnam. *Working paper*.

Tenorio, R. and T. Cason. 2002. To Spin or Not To Spin? Natural and Laboratory Experiments from The Price is Right. *Economic Journal* 112: 170-195.

Urzua, S. and J. J. Heckman. 2009. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. In *UCD Geary Institute Discussion Paper Series* University College Dublin.

Webb, E., D. Campbell, R. Schwartz, and L. Sechrest. 1999. *Unobtrusive Measures: Revised Edition* SAGE Publications, Inc.

Wooders, J. 2010. Does Experience Teach? Professionals and Minimax Play in the Lab. *Econometrica* 78(3): 1143-1154.