What Do Laboratory Experiments Tell Us About the Real World?

John A. List
University of Chicago and NBER

and

Steven D. Levitt*
University of Chicago and American Bar Foundation

20 September 2005

Abstract

Perhaps the most important question facing experimental economists is whether, and to what extent, behavior in the laboratory is a good indicator of behavior in the field. Lab environments differ from naturally occurring environments along numerous dimensions: subjects know that they are being scrutinized, special emphasis is placed on the process by which decisions are made and final allocations are reached, the stakes are typically small, and the participants are self-selected. In this paper, we develop a simple model that incorporates these various factors. Combining the predictions of the model with a review of the empirical evidence from both inside and outside the lab, we argue that experimenter scrutiny and the emphasis on process pose the greatest difficulties for extrapolation of experimental results measuring pro-social behaviors to the real world. We find less evidence that low stakes distort behavior in such experiments, but stakes can influence tasks that involve high computational costs. Self-selection of participants, likewise, appears to be of lesser importance for experiments measuring social preferences, although for many other experimental games an understanding of selection, context, and external market forces is necessary for proper inference. Furthermore, it is possible that previous experimental findings have exaggerated behavioral differences across important dimensions such as gender and race.

JEL: C9 (Design of Experiments), C93 (Field Experiments)

Key words: generalizability; laboratory experiment; field experiment

*Thanks to seminar participants at the 2005 International Meetings of the ESA for useful suggestions. Suggestions from Glenn Harrison, Jayson Lusk, Charles Plott, and Robert Slonim improved the study. Sherman Shapiro and the National Science Foundation provided financial support.

Nearly 400 years ago, Galileo performed the first recorded laboratory experiment, timing balls as they rolled down an inclined plane to test his theory of acceleration (Settle, 1961). Since that time, laboratory experiments have been a cornerstone of the scientific method. Feynman (1963) illustrates this fact when noting that "The principle of science, the definition almost, is the following: The test of all knowledge is experiment. Experiment is the sole judge of scientific 'truth.'"

A critical maintained assumption underlying laboratory experiments is that the insights gained in the lab can be extrapolated to the world beyond, a principle we denote as generalizability.[1] For physical laws and processes (e.g. gravity, photosynthesis, mitosis), the evidence to date supports the idea that what happens in the lab is equally valid in the broader world. Shapley (1964, p. 43), for instance, noted that "as far as we can tell, the same physical laws prevail everywhere." Likewise, Newton (1687; p. 398 (1966)) scribed that "the qualities....which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever."[2]

Increasingly, economists have been using the traditional experimental model of the physical sciences as a method to understand human behavior. Holt (2005) documents that

[1] Many different types of phrases have been used to depict the relationship between the lab and the field, with "parallelism," "external validity," and "ecological validity" being the most popular. Parallelism, which we traced to Shapley (1964), is said to be established if the results found in the laboratory hold in other, particularly real-world, situations under *ceteris paribus* conditions (see, e.g., Wilde, 1981; Smith, 1982). Campbell and Stanley (1963) introduced external validity as follows: "external validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (p. 5). Ecological validity has taken on a multifarious set of meanings, including the notion that a study is ecologically valid if "one can generalize from observed behavior in the laboratory to natural behavior in the world" (Schmuckler, 2001, p. 419). But, confusion arises because it is clear that Egon Brunswik coined the term *ecological validity* to indicate the degree of correlation between a proximal (e.g., retinal) cue and the distal (e.g., object) variable to which it is related (see, e.g., Brunswik, 1955, 1956).

[2] The importance of this metaphysical principle should not be underestimated. Consider the recent revolution of understanding concerning the planet Pluto. When James Christy and Robert Harrington discovered Pluto's moon, Charon in 1978, planetary scientists hailed the discovery because they could now calculate Pluto's mass accurately by using the orbiting period and the laws of gravitation. Measurements subsequently informed planetary scientists that the Pluto-Charon system is about 1/400th the mass of earth, much smaller than the first estimate that Pluto was about 10 times as massive as the earth (Binzel, 1990).

experimental economics is a "boom industry," showing that publications using the methodology were almost non-existent until the mid-1960s, surpassed 50 annually for the first time in 1982, and by 1998 there were more than 200 experimental papers published per year.

This explosion of research in experimental economics has had an important influence on the profession. For example, in the case of modeling preferences, a series of experiments revealing a disparity between subjects' willingness to pay and willingness to accept, has spurred an entire body of theoretical research (e.g., Kahneman and Tversky, 1979; Loomes and Sugden, 1982; Heiner, 1983; Hanemann, 1991; Tversky and Kahneman, 1991). Likewise, the finding that subjects seemingly have interdependent utilities in various environments has led to a theoretical exploration of the economic consequences of social preferences (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002; Charness and Rabin, 2002). Indeed, the foundations of behavioral economics are firmly anchored in the observation that human decision making in the laboratory often deviates from the predictions of standard economic models (Rabin, 1998; Mullainathan and Thaler, 2000). Another indication of the impact of experimental economics on the discipline is Vernon Smith's 2002 Nobel Prize in Economics for his pioneering work in the field.[3]

The allure of the laboratory experimental method in economics is that, in principle, it provides *ceteris paribus* observations of motivated individual economic agents, which are otherwise exceptionally difficult to obtain using conventional econometric techniques. Lab experiments provide the investigator with a means to directly influence the set of prices, budget sets, and actions available to actors, and thus measure the impact of these factors on behavior within the context of the laboratory.

[3] The Nobel Prize Committee described Smith's contribution as "having established laboratory experiments as a tool in empirical economic analysis, especially in the study of alternative market mechanisms" (2002 Nobel Prize Announcement).

The basic strategy underlying laboratory experiments in the physical sciences and economics is similar, but the fact that humans are the object of study in the latter raises fundamental questions about the ability to extrapolate experimental findings beyond the lab that do not arise in the physical sciences.[4] While few would question whether $Uranium_{239}$ would emit beta particles and turn into Neptunium in the presence or absence of scientists, human behavior may not be governed by immutable laws of nature. In particular, we emphasize four characteristics of laboratory economic experiments that raise potential problems in generalizing results from the lab to the outside world:[5]

1) *Subjects know they are being watched.* Humans, unlike Galileo's rolling balls or $Uranium_{239}$, know that they are participating in experiments. Making decisions in an artificial environment and general awareness of the fact that their actions are being observed and recorded might influence how people behave. This notion is perhaps best illustrated by a nearly century old quip of A.H. Pierce (1908):

> It is to the highest degree probable that the subject['s] . . . general attitude of mind is that of ready complacency and cheerful willingness to assist the investigator in every possible way by reporting to him those very things which he is most eager to find, and that the very questions of the experimenter . . . suggest the shade of reply expected .... Indeed . . . it seems too often as if the subject were now regarded as a stupid automaton

Decades of research within psychology highlight the power of the experimenter within the experimental situation (Orne 1959a, 1959b, 1962, Rosenthal 1967, 1976, 1991,

[4] We take as given that the issue of generalizability is important. There are instances where generalizability might not be of first rate importance, however. For example, Mook (1983) and Schram (2005) detail some arguments of cases where they believe that it is not important. Schram (2005) notes that the "classic" view of experimentation is one of theory-testing, arguing that "external validity is not an issue" in these instances. Writings of Plott (e.g., 1982) and Smith (e.g., 1982) reinforce this point. Another example includes using the lab for methodological purposes—i.e., to inform field designs by abstracting from naturally-occurring confounds.

[5] Taking a different approach to the same problem, Cross (1980, p. 405) argues that "it seems to be extraordinarily optimistic to assume that behavior in an artificially constructed "market" game would provide direct insight into actual market behavior." Cross provides four problems he has with experiments: i) subjects might view themselves in a "game" against the experimenter, ii) subjects might be role-playing, iii) real world behavior is typically learned over months and years whereas experiments are short term affairs, and iv) subjects bring context into the lab.

Milgram 1963, Haney et al. 1973).[6] The strength of such factors is well-understood in the social psychology literature, where experimenters commonly deceive subjects to control for demand-induced effects (Kagel et al., 1979).

2) *Context matters and it is not completely controlled by the experimenter.* Human behavior is heavily influenced by context. The actions humans take are influenced by a dazzlingly complex set of relational situations, social norms, past experiences, and the lessons gleaned from those experiences. Consequently, the experimental investigator does not have complete control over the full context within which the subject makes decisions. In experimental settings, it has been shown that subtle manipulations in the dictator and ultimatum games can have drastic effects on play (Hoffman et al., 1994); likewise, Roth (1995, p. 282) showed that a laboratory experiment framed in a market context generates choices that are more in conformance with self-interest than do other non-market environments. Furthermore, rates of defection in prisoner dilemma games swing wildly depending on if subjects are playing a "Community" or "Wall Street" game (Ross and Ward, 1996); more generally, using terms like "opponents" versus "partners," asking people to "contribute" or to "allocate" funds to a public good, or to "punish" or to "assign" points" to other agents in a punishment game all influence play (Gintis, 2001). Contextual factors that are beyond the control of the experimenter may have equally profound impacts on actions.

[6] For instance, Orne (1962) writes "Just about any request which could conceivably be asked of the subject by a reputable investigator is legitimized by the quasi-magical phrase, "This is an experiment," and the shared assumption that a legitimate purpose will be served by the subject's behavior." Relatedly, in physics the Heisenberg Uncertainty Principle reminds us that the act of measurement and observation alters that which is being measured and observed. And, the Hawthorne Effect is a famous example of an instance where people changed their behavior merely because they were being watched. The fact that some agents alter their behavior when observed by others has also been termed the "interpersonal self-fulfilling prophecy" and the "Pygmalion Effect" (see Harrison and List, 2004).

3) *Small* Stakes. The stakes for which experiments are played are typically quite small,[7] but researchers often wish to extrapolate the behavioral implications to real-world settings in which the decisions are far more important.

4) *Self-Selection.* Human subjects cannot generally be compelled to participate in experiments, but rather, self-select into them. Those observed in experiments may systematically differ from the relevant population to which the results will be extrapolated. For instance, those who select into laboratory experiments are generally volunteers interested in the research project, more likely to be "scientific do-gooders," or students who readily cooperate with the experimenter and seek social approval (Orne, 1962; Rosenthal and Rosnow, 1969).[8] In contrast, market participants are likely to be a highly selected sample of individuals whose traits allow them to excel in the marketplace. Individuals suffering from decision-making defects will tend either to adapt, to hire agents who do not have such defects to represent them, or simply to disappear from the market.

The remainder of this paper is devoted to understanding the degree to which these four elements of experimentation with human subjects interfere with generalization of lab results to the broader world, the sorts of questions for which these problems will be most or least severe, and the steps that researchers might take to minimize these distortions.[9]

---

[7] One of the major differences between economic experiments and psychology experiments is that economists use monetary rewards that are contingent on behavior whereas psychologists typically rely on hypothetical incentive schemes. A related issue is whether subjects behave differently over "house" money and own-(or earned) money. In simple dictator games, empirical evidence suggests that dictators give less when they are allocating earned money or have earned their position in the game (see, e.g., Hoffman et al., 1994; List and Cherry, 2005).

[8] For example, when experimentally naïve high school students were asked "How do you think the typical human subject is expected to behave in a psychology experiment?" over 70 percent circled characteristics labeled cooperative and alert (Rosenthal and Rosnow, 1973, pp. 136-137).

[9] The list we present certainly does not exhaust the set of reasons that lab experiments may not provide direct guidance with respect to behavior outside the lab. For instance, subjects tend to have less experience with the games they play in the lab relative to situations that frequently arise in naturally occurring settings. Moreover, experiments

We begin Section I by presenting a simple model that provides a convenient framework for understanding how, and why, these issues are important. The model stresses the fact that utility maximization is driven not only by wealth maximization, but potentially also by an individual's desire to "do the right thing," or make the "moral" choice. The weight that an individual places on "doing the right thing" is likely to increase when a subject is being watched, or the process by which choices and final allocations are reached is emphasized. Self-selection of individuals may be important if the weights given to moral versus wealth-maximizing considerations vary across individuals faced with the same problem, or people disagree as to what the "correct" moral choice is in certain situations. The stakes of the game, in our model, have an ambiguous effect on behavior because they potentially affect both the determination of what is the moral action and the relative weights given to morality versus wealth.

Besides having practical predictive importance for interpreting laboratory evidence, the model also sheds light on behaviors outside of the lab that are sometimes seen as puzzling, for instance why diners tip at restaurants in which they never intend to return, and why out-of-state automobile drivers provide a fellow motorist entering the highway ample space to merge even though they themselves plan to exit immediately.

In sections II-IV we illustrate the usefulness of our model by exploring its implications for interpreting the results in three quite distinct strands of the experimental literature: (1) studies measuring pro-social preferences, (2) games in which "doing the right thing" is not an issue, and (3) experiments designed to measure differences in behavior across broad groups, e.g., by gender

---

typically have short durations (minutes or hours), whereas many real life decisions (e.g. effort to exert in labor market settings) are made over much longer timeframes. We only briefly touch on these issues below. Other arguments exist as well (see, e.g., Kagel et al., (1979), Cross (1980), Starmer (1999a, 1999b), and Hertwig and Ortmann (2001)). Interestingly, for a relatively new field that is developing quickly, we find it surprising that little methodological discourse has occurred in experimental economics. Perhaps this is not surprising in light of Samuelson's (1963, p. 231) tongue-in-cheek remark some four decades ago: "Methodological discussion, like spinach and calisthenics, is good for us..." We hope that our study will have some small influence in motivating further discussion in this area.

or race. We argue that the combination of being watched and an emphasis on process lead to over-estimates of pro-social behaviors. This hypothesis is supported by the results from field experiments and non-experimental studies. In situations where morality and wealth are not competing objectives, such as bidding in common value auctions, we find evidence that individuals suffering from decision-making defects (such as the Winner's Curse) tend either to adapt, disappear from the market, never enter the market at all, or participate in markets that have developed securities to protect bidders from falling prey to such defects. Laboratory experiments may also provide misleading conclusions about differences in behavior across groups, such as men and women, although our evidence on this is more speculative.

Combining the theoretical model and the data, we argue for an important, but inherently limited, role for laboratory experiments in economics. In isolating a single factor for investigation, laboratory experiments necessarily create an environment that is not representative of "real life" situations. The creation of a very stylized environment allows for causal inference under the precise experimental conditions. This role is useful in confirming a cause-effect relationship scholars strive for in theory testing (commonly termed "internal validity"). Our findings, however, indicate that laboratory analysis alone informs us little about whether, and to what extent, this situation manifests itself in the real world and how important the effect is; rather it tells us what *can* happen and provides an underlying mechanism that might be at work when certain data patterns are observed in the field. This does not mean that laboratory experiments are not valuable, only that blindly extrapolating lab findings to the real world is ill-advised. When thinking about what a particular laboratory experiment implies for actions outside the lab, a sensible theoretical framework for understanding the potential distortions to behavior induced by the lab is a necessary ingredient.

Section V concludes with a summary and draws out some implications for future research in experimental economics. One fruitful avenue to form a bridge between lab experiments and reality is field experiments. Our belief is that proper utilization of field experiments can mark the beginning of the end of "schools" of thought in empirical economics. By carefully exploring theoretically and empirically the nature and extent of the various factors that potentially influence lab behavior, experimentalists and non-experimentalists alike can begin to discuss issues of the day using similar logic.

I. Model

We begin by developing a model that makes precise our arguments regarding the potential factors that might interfere with a seamless extrapolation of behavior from the laboratory to naturally-occurring settings. In building the model, we err on the side of simplicity, sacrificing generality in the process. Nonetheless, we believe that the basic insights of the model will hold in a much more general framework.

A utility-maximizing individual i is faced with a choice regarding a single action $a \in (0,1)$. The choice of action affects the agent's utility through two channels. The first effect is on the individual's wealth (denoted W_i). The higher the stakes or monetary value of the game, which we denote v, the greater the decision's impact on W_i. In most experimental circumstances we have in mind, the wealth maximizing action is taken as the default prediction of economic theory, for example, playing the dominant strategy of zero contributions in a one-shot public goods game in which the private return exceeds the public return.

Decisions can have an impact on individual utility that goes beyond changes in wealth (Smith, 1759; Becker, 1974).[10] If, for instance, an individual is altruistic, he will derive utility

[10] Smith viewed decisions as a struggle between "passions" and an "impartial spectator," who was a "moral hector who, looking over the shoulder of the economic man, scrutinizes every move he makes" (Grampp, 1948, p. 317).

from charitable contributions. More generally, we have in mind that decisions which an individual views as immoral, anti-social, or at odds with his or her own identity (Akerlof and Kranton 2000; 2005) may impose important costs on the decision maker. For instance, in a dictator game, keeping a greater share for oneself increases an individual's wealth, but being selfish may cause the agent moral disutility. Formally, we denote as M_i the non-pecuniary moral cost or benefit associated with action i. This moral payoff might vary across people, religions, or societies, but examples of the types of choices that might maximize M_i would be the Golden Rule of "Do unto others as you would be done by" or Socrates' moral code "Do not do to others what would anger you if done to you by others."[11]

In practice, many factors influence the moral costs associated with an action across people, contexts, and societies, but for modeling purposes, we focus on just two aspects of the moral determinant. The first of these is the financial externality that an action imposes on others. The greater is the negative impact of an action on others, the more negative the moral payoff M_i.[12] We model the externality as being a function of the stakes of the game v. The second factor that influences the moral choice is the set of social norms or legal rules that govern behavior in a particular society. For instance, extending one's middle finger vertically while the other fingers are in a fist is deemed extremely rude in American culture, but not in many other parts of the world. Although there is no financial externality borne by the recipient, there

---

For further discussion on this issue see the entertaining discussion of "Adam Smith, Behavioral Economist," in Ashraf et al. (2005). Becker (1974) introduced altruism as an explanation for why individuals give (or cooperate) when it is money-maximizing not to give (or cooperate).

[11] Formalizing the notion of morality in economic models is not novel to this study. For example, in the charitable giving literature, an important alternative modeling approach to the pure altruism framework is to consider moral or group interested behavior (see, e.g., Laffont (1975), Sen (1977), Collard (1978), Margolis (1982), and Sugden (1984)). In Sugden, (1984), for instance, agents adhere to a "moral constraint," whereby they compare themselves to the least generous person when making their contributions. Moral concerns are also present in the work of Frey et al. (1996), Karni (1998), and Kotchen (2005), for example. Relatedly, in Bernheim's (1994) conformity model agents care about their status, and behavioral departures from the social norm impairs status. Akerlof (1982) and Jones (1984) obtain similar conformity results by assuming deviations from social norms have direct utility consequences.

[12] In Andreoni (1989; 1990), purely egoistic models assume that agents are not concerned with final outcomes.

nonetheless are potential moral and social costs associated with such behavior. Likewise, the mere fact that an action is illegal (e.g., illicit drug use or smoking in restaurants), may impose an additional cost for partaking in such behavior. Mathematically, we denote these social norms against an action as *n*, with a greater value of *n* associated with a stronger norm against a behavior.

For expositional purposes, we adopt a simple utility function in which wealth and moral considerations enter in an additively-separable form:

$$U_i(a,v,n) = \theta_i M_i(a,v,n) + (1-\theta_i)W_i(a,v) \qquad (1)$$

where θ reflects the relative weight that individual *i* places on "morality" versus wealth, *M* and *W* are the respective moral and wealth payoffs, *a* is the action taken, *v* reflects the stakes of the game, and *n* are the social norms against an action. The moral and wealth payoffs of a particular action may vary across individuals, as may the relative weight that is assigned to those two factors.[13]

The factors that influence the weight θ that an individual places on moral versus financial considerations are critical to the model. The more closely an individual's actions are being scrutinized (e.g., the act is being televised, or is taking place in front of one's children, or a respected figure such as a member of the clergy), moral concerns are likely to receive increased emphasis. And, experimenter "demand effects" teach us that the greater the emphasis on the process by which a decision is reached, the more important are moral concerns. For instance, a natural tendency observed by social psychologists is for subjects to infer from elaborate explanations that they are supposed to behave in a certain way (Loewenstein, 1999). Denoting

[13] In the model social norms against an action do not have any direct financial impact, although one could imagine a generalization of the model involving repeated play in which such a financial impact would be present, e.g., if violators of social norms were punished by non-cooperation in the future.

these two aspects of scrutiny and process-emphasis as *s*, and noting that the only choice variable for the individual is the action *a,* the decision problem faced by the individual is

$$\underset{a\in(0,1)}{MAX}\ U_i(a,v,n,s) = \theta_i(s) * M_i(a,v,n) + (1-\theta_i(s)) * W_i(a,v) \qquad (2)$$

Solving this simple decision problem yields several predictions. First, when the wealth maximizing action has a moral cost associated with it, the agent will (weakly) deviate from that action towards one that imposes a lower moral cost. The greater is the social norm against the wealth maximizing choice, the larger the deviation from that choice. Second, the greater is θ (the weight given to moral considerations), the more the agent deviates from the wealth maximizing choice. Moral considerations are likely to be most important when an agent's actions are being closely scrutinized and the process by which decisions and allocations are reached is emphasized. Third, when the moral and wealth maximizing actions differ, increasing the stakes of the game has an ambiguous effect on the action taken. Raising the stakes increases the value of pursuing the wealth-maximizing action, but the moral cost associated with that choice will also rise. Either effect can dominate. Fourth, to the extent that individuals follow different moral codes (that is, $M_i \neq M_j$ for individuals *i* and *j*), or place different weights on moral versus wealth considerations ($\theta_i \neq \theta_j$), they will generally make different choices when faced with the same decision problem.

It is useful to consider the predictions of the model in terms of the dimensions along which we believe laboratory environments may systematically differ from naturally-occurring environments. Most importantly, the emphasis on process and experimenter scrutiny might exaggerate the weight placed on moral concerns in the lab relative to the outside world. In situations where moral and wealth maximizing choices are not aligned, therefore, we expect that behavior in the lab will deviate more from the wealth maximizing choice than behavior in

naturally-occurring situations, such as anonymous market transactions.[14] The potential bias associated with the small stakes typically used in the lab, according to our model, has an ambiguous effect on actions. Self-selection of participants into laboratory experiments may affect the actions observed in the lab; the direction in which behavior is distorted depends upon the nature of the selection.

In the next sections, we focus on three different strands of the experimental literature, comparing the predictions of the model with the existing evidence from experimentally generated lab and field data as well as non-experimental data.[15]

II. Implications of the model for games on the importance of social preferences

One of the most influential areas of research in experimental economics in recent years has been on games that provide insights into social preferences.[16] This broad class of games includes bargaining games (Roth, 1995), public goods games (Ledyard, 1995), and trust, or gift exchange, games (e.g., Camerer and Weigelt, 1988; Fehr et al., 1993; Berg et al., 1995; Charness, 1996; Fehr et al., 1997; Fehr and Falk, 1999; Fehr and Gachter, 2000; Gächter and Falk, 2002; Hannan et al., 2004; Brown et al., 2004). Findings from gift exchange games, for example, have been interpreted as providing strong evidence that many agents behave in a reciprocal manner even when the behavior is costly and yields neither present nor future material

[14] Additionally, moral issues might play a larger role in the lab than in the marketplace because expressing moral views might seem "less costly" in the lab since subjects are playing over "house" money.

[15] Our model also lends insights into policymaking. For example, officials interested in benefit/cost analysis might find our approach convenient in explaining why economic values from contingent valuation exercises contain hypothetical bias (see, e.g., Diamond and Hausman, 1989) and are upwardly biased due to social influence effects (see, e.g., List et al., 2004). More generally, if policymakers are able to manipulate the relative weights the population places on wealth and morality, then certain outcomes are more likely to be achieved. One method to achieve such goals is to inform children and adults about the dangers of certain activities. This is consonant with certain public programs such as "Just say no (DARE)" and "Practice Safe Sex," which are initiatives that teach people about the dangers of drugs and unprotected sex in hopes of inoculating them to "just say no" to such activities.

[16] While various definitions exist, we define an agent with social preferences as one who has preferences that are measured over her own and others' material payoffs. Such preferences might arise due to, for example, altruism, reciprocity, fairness, and inequality-aversion.

rewards. Further, the results have been widely applied outside the laboratory, based on the assumption that the experimental findings are equally descriptive of the world at large. For instance, Fehr et al. (1993, p. 437) note that their results "provide…experimental support for the fair wage-effort theory of involuntary unemployment" (see also Camerer, 2003; Sobel, 2002).[17]

Viewed through the lens of our model, however, there are a number of reasons why laboratory studies purporting to measure social preferences may be a poor guide to behavior outside the lab.[18] In what follows, we examine the empirical evidence regarding each of the possible complications to extrapolating the experimental findings outside the lab highlighted in our model. We conclude that the fact that subjects know that they are being watched is a fundamental obstacle to extending lab findings on social preferences to the world at large, but the other factors are of less consequence.

Subjects know there are spectators in the lab

In the typical lab experiment, subjects are fully aware that their behavior is being monitored, recorded, and subsequently scrutinized, possibly inducing them to behave in a way that portrays them in a positive light or pleases the experimenter. Since there is a divergence between the moral choice and the wealth-maximizing choice in these games, the weight placed on acting morally will likely be greater in the lab because players know that they are being

[17] Methodologically, this approach has yielded a fundamental alteration in the manner in which experimental economists view the nature of the laboratory, effectively turning the laboratory into a venue for *measuring* individual preferences, rather than ascertaining how such preferences affect market outcomes.

[18] We anticipate that at this point our critics will contend that "these experiments merely show what *can* happen, not what will necessarily happen in all economic domains." We have two responses. First, then such laboratory experiments are not that enlightening because a thought experiment can easily yield similar insights on what *can* happen in the laboratory in these types of games. Second, this argument is not consistent with the manner in which the received results have been interpreted in the literature (see any of the gift exchange studies cited above).

watched. The amount of unselfish behavior will be higher in the lab than in the real world situation modeled by the experimenter.[19]

A prediction of our model is that taking steps to reduce the extent to which subjects are knowingly observed should reduce the amount of pro-social behavior. One way of accomplishing this is to create an environment that provides anonymity for the subjects in the laboratory. Hoffman et al. (1994; 1996), for instance, used a "double-blind" approach whereby the experimental monitor could not infer individual subjects' actions in a simple dictator game. In the same spirit, List et al. (2004) used a "randomized response" technique to ensure that subjects would preserve their anonymity when asked to contribute money to a real public good. Both studies, and other related research (Eckel and Grossman, 1996; Masclet et al., 2003; Rege and Telle, 2004), find greatly reduced giving. Hoffman et al. (1994), for example, find that 22 of 48 dictators (46%) donate at least $3 of a $10 pie under normal experimental conditions, but when subject-experimenter anonymity is added, only 12 of 77 dictators (16%) give at least $3. Hoffman et al. (1994, p. 371) conclude that observed "behavior may be due not to a taste for "fairness" (other-regarding preferences), but rather to a social concern for what others may think, and for being held in high regard by others."[20]

The experimental approaches described above, however, are subject to criticism. For example, lessons learned from social psychologists teach us that such efforts to ensure anonymity might result in subjects inferring that the experimenter "demands" them to behave in

[19] Of course, agents are oftentimes observed in many naturally-occurring environments as well. The point here is that in cases where the experimenter is modeling behavior in one-shot, anonymous settings, for example, the lab might not provide an environment conducive to such observation. Our theory highlights the importance of implementing an experiment with the same level of oversight that corresponds to the field situation one is attempting to model. If this chore is not possible, then an appropriate model of decision-making should be considered.

[20] Relatedly, our model also provides an explanation for the experimental results that show declining positive survey responses in favor of a proposed public project as the likelihood that the survey is executed 'for real' increases (see, e.g., Cummings and Taylor, 1998).

a manner that might be deemed unacceptable (Loewenstein, 1999). Alternatively, subjects in such environments might not fully embrace the anonymity promise: informally we polled subjects who were part of a recent anonymous experiment on whether they believed it was truly anonymous and a non-trivial portion of subjects reported "that the experimenter could likely determine identities." Thus, such important differences in behavior observed across anonymous and non-anonymous settings might be a lower bound estimate of the true spectator effect predicted by our model.

An alternative approach to achieving participant anonymity is to move from a laboratory environment to a naturally occurring environment in which the actors are unaware that their behavior is being monitored, as in List's (2005) analysis of the sportscard market. In particular, List carried out a natural field experiment version of the gift exchange game: he sent confederates as buying agents to approach dealers on the floor of a sportscard show, instructing them to offer various different prices in return for requests of differing quality levels. Since quality is difficult to detect in this market for untrained consumers, like the data from the mirror laboratory experiments that List executed, if social preferences play a role in this case the card's grade and the price offer should be positively correlated. Once the buying agents had purchased each of the cards from the dealers, List had every card professionally graded. The results were again consistent with our model. In cases when the dealer believed that consumers could not have the card graded or when there was likely to be little future interaction, no relationship between price and quality emerged. In other words, the social preferences so routinely observed in the lab completely disappeared, even though the same agents revealed strong evidence of having social preferences in laboratory experiments.

Other field generated data lead to similar conclusions. For example, making use of personnel data from a leading United Kingdom based fruit farm, Bandiera et al. (2005) find that behavior is consistent with a model of social preferences when workers can be monitored, but when workers cannot be monitored, pro-social behaviors disappear. Being watched proves to be the critical factor influencing behavior in this study.

Finally, Benz and Meier (2005) combine insights gained from a controlled laboratory experiment and a natural field experiment to compare how individuals behave in donation laboratory experiments and how the same individuals behave in a naturally-occurring charitable donation situation. They find some limited evidence of similar behavior, but ultimately conclude that "the correlation found between experimental and real-life behavior is rather weak."

Context matters and it is not completely controlled by the experimenter

Experimentalists are fully aware that context in their instructions, inducing "role" playing, framing, and the like can influence subject behavior (see, e.g., Hertwig and Ortmann, 2001). Accordingly, our model predicts that even in games with extreme perfect equilibria, theoretically-consistent behavior can be observed since subjects will use the contextual cues of the game to figure out which set of norms applies to the particular problem at hand.[21] One subtle aspect of the context that the experimenter can control that has not been fully explored is the fact that restrictions on the available choice set affect behavior. Consider a dictator game, where one agent unilaterally determines how a prize should be split between himself and another player, and that decision determines the allocation. In the typical dictator game, the subject is given, say, $10 and asked what portion he would like to share with the other player. The experiment is framed such that "giving nothing" is the least generous act, and substantial sums of money are

[21] The literature on "best-shot" and "impunity" games are examples (see Roth, 1995).

given away. If instead, the subject is given a $10 prize and is told that he can give any portion of his prize away to the second player, or confiscate up to an additional $10 from the other player, subjects give little to the other player.[22] Importantly, in real-world contexts, there is typically the option of both giving and receiving, which may help explain in part why, contrary to the lab environment, people rarely give generously to random strangers on the street or train.

Important aspects of context are also not under the control of the experimenter. For instance, lab experiments testing for pro-social behavior are designed to be "one shot" affairs: the subjects enter the lab, play various games over a finite horizon, and then depart with cash earnings in hand. As various authors have pointed out (e.g., Hoffman et al., 1996; Ortmann and Hertwig, 2000; Harrison and Rutstrom, 2001), public good provision, dictator, ultimatum, trust, and gift exchange games are typically not one-time encounters, but rather repeated games, in the outside world. To the extent that participants perceive these games in the laboratory as some form of social dilemma, they are likely to retrieve experiences and strategies that, unbeknownst to the experimenter, change the nature of the games (see also Binmore 1994). Effectively, personal experiences may cause the subjects to play these one-shot games as if they are repeated games, and the experimenter may have little or no ability to moderate this phenomenon.[23] Consider, for instance, the case of sportscard markets. List (2005) finds that in this naturally occurring setting, when the good could be easily graded and the dealer would likely have future interactions, sellers provide higher quality cards to those buyers who offer more money. This

[22] This effect is observed in data that we gathered as well as in Bardsley (2005), who uses a much different set of treatments.

[23] There is mixed evidence on the ability of students to demonstrate clear strategic adjustments as they move from one-shot to repeated contexts. Whereas Gaechter and Falk (2002) argue that they can perceive a difference in gift exchange games, Roth et al. (1991) find that it does not occur in ultimatum games.

result, presumably driven by the desire of sellers to build a favorable reputation, would be indistinguishable from pro-social preferences in a lab environment.[24]

Perhaps the most telling evidence regarding the importance of context comes from Henrich et al. (2001; 2004). In the latter study the formidable group of scholars conducted ultimatum, dictator, and public goods games in fifteen different small-scale communities in developing countries. They found enormous variation in behavior across communities, differences they are able to relate to interactional patterns of everyday life and the social norms operating in these various communities. For instance, as Henrich et al. (2004, p.31) note, the Orma readily recognized "that the public goods game was similar to the *harambee*, a locally-initiated contribution that Orma households make when a community decides to construct a public good such as a road or school," and subsequently gave quite generously. Likewise, among the whale hunting Lamalera of Indonesia and the Ache in Paraguay, societies with strong norms of sharing, very generous ultimatum game offers are observed and very few offers are rejected. Alternatively, in small-scale foraging societies that are characterized by a much different ethnographic description, such as the Hadza of Tanzania, low offers and high rejection rates are observed in ultimatum games. As Henrich et al. note (2004, p. 33), these "contrasting behaviors seem to reflect their differing patterns of life, not any underlying logic of hunter-gatherer life ways."

[24] List (2005) also provides a different form of evidence on the role of context. List began by having sportscard traders play standard gift exchange games in a laboratory setting. He then undertook a series of treatments where he successively added context that approached the sportscard trading environment. He began by simply using terminology that matched the language of the marketplace and concluded with an experimental lab market where buyers and sellers exchanged cash for goods of uncertain quality in face-to-face transactions. Except for the fact that the final experiment in this series was conducted in a laboratory setting, the game structure mirrors the actual decision-making process in the marketplace from which these subjects were drawn. Only minor differences in behavior were observed as List varied the context in this way, suggesting that it was not the unfamiliarity of the standard gift exchange game that was driving evidence consistent with pro-social behavior.

Critically, in all of the experiments Heinrich et al. (2001; 2004) conducted, the context that the experimenter can control (i.e. the payoffs, the description of the way the game is played, etc.) was almost identical. It was the part of the context that actors themselves bring to the game and experimenters cannot control (e.g. past experiences and internalized social norms) that proved centrally important in the outcome of play.

Another related contextual concern is whether the games played in laboratory experiments are reflective of the real life situations they attempt to model and to which inference is ultimately made by the experimenter. For instance, in a recent study, Eckel and Wilson (2004), asked subjects who just participated in a trust game "what does this situation remind you of?" Only 29/110 (10/108) of first (second) mover subjects mentioned trust. Of course, if the participant is playing a different game than the experimenter believes, then any inference is potentially dubious.

Small stakes

Our model predicts that in games that have both a morality and wealth component, the stakes effect is ambiguous since both M and W are increasing functions of v. The evidence in the literature is roughly consonant with this view. For example, in dictator and ultimatum games the influence of stakes has been negligible on proposers' behavior (see, e.g., Slonim and Roth, 1998: Cameron, 1999; Forsythe et al., 1994; List and Cherry, 2005; Carpenter et al., 2005). In terms of trust and gift exchange games, Fehr et al. (2002) report that fairness concerns play an important role for both low and high stakes games whereas Parco et al. (2002) find that raising financial incentives causes a breakdown in mutual trust in centipede games.

Self-selection into the experimental subject pool

Most experiments measuring social preferences have been conducted using students who self-select into the experiments. To the extent that these students have different weights on moral versus pecuniary factors, or different moral codes, laboratory findings may not provide accurate guidance for behavior in naturally-occurring situations for which the experimenter is attempting to make inference.

One approach to investigating biases from this source is to examine whether professionals, or other representative agents, and students behave similarly in laboratory experiments.[25] In order for these laboratory findings to be meaningful, however, it must be the case that the extent of lab-induced changes in behavior (e.g., the increase in emphasis on the moral action) is the same across both groups. Fehr and List (2004) examine experimentally how Chief Executive Officers (CEOs) in Costa Rica behave in trust games and compare their behavior with that of Costa Rican students. They find that CEOs are considerably more trusting and exhibit more trustworthiness than students.[26] These differences in behavior may mean that CEOs are more trusting in everyday life, or it may be that CEOs are more sensitive to the lab's scrutiny, or that the stakes are so low for the CEOs that the sacrifice to wealth of making the moral choice is infinitesimal.

Even if strong insights could be gained about subject pool differences from these artefactual field experiments, a related issue concerns the possibility that only certain types of participants—students or professionals—select to take part in the experiment. In this case, if the

[25] Harrison and List (2004) denote this type of exercise as an "artefactual" field experiment, in that it is identical to a typical laboratory experiment but one which makes use of a non-standard subject pool. See http://www.arec.umd.edu/fieldexperiments/ for a bibliography of nearly 100 artefactual field experiments.

[26] Consistent with Fehr and List (2004), in dictator games, Carpenter et al. (2005) find that employees at a Kansas City distribution center are more generous than students. Harbaugh et al. (2003), conducted a set of trust experiments with students in third, sixth, ninth, and twelfth grade and found little variation across the participants in terms of trust and trustworthiness. Alternatively, in dictator games, the youngest children tend to make considerably smaller transfers than do older children and adults in Harbaugh et al. (2003).

selection rule differs across subject pools then valid inference might be frustrated. For example, volunteers who have social preferences or who readily cooperate with the experimenter and seek social approval might be those who are most likely to participate in the experiment. In this case, games that purport to measure pro-social behaviors will yield upper bound estimates on the propensities of the target population.

There exists some limited, but suggestive data from field and lab experiments that provide some support for this argument about selection into laboratory gift exchange experiments. List (2005) approached a number of sportscard sellers about participating in a laboratory experiment. Some sellers declined his invitation, but later and unbeknownst to them, participated as sellers in a gift exchange field experiment. Those who declined to participate in the lab portion of the experiment were less pro-social in the field compared to dealers who agreed to participate in the lab experiment, although the differences are not statistically significant. Likewise, Armin Falk and Stephan Meier explore the types of agents who participated in Ernst Fehr's gift exchange studies at the University of Zurich and found that the subjects were not a random sample of students. Those who tended to contribute more to student charitable funds were more likely to participate in Fehr's experiments, suggesting that the subject pool is biased towards pro-social agents.[27]

III. Implications of our model for experiments in which the moral and wealth-maximizing actions are not competing objectives

In Section II we emphasized how the lack of congruence between moral and wealth-maximizing actions can lead laboratory experiments to yield findings that may not be readily extrapolated to the outside world. In the large class of games where there is no inherent conflict between the moral choice and the wealth-maximizing choice (e.g. experiments exploring

[27] Personal correspondence, Stephan Meier.

Bayesian updating, risk and uncertainty, psychological biases such as loss aversion, hyperbolic discounting, auctions, market experiments where the demand and cost functions are unknown, lobbying games, etc.), many of the concerns we have emphasized become inconsequential. In our model, when $M=W$, the weight placed on each becomes irrelevant, and the maximization problem in equation (1) becomes straightforward. The only potential distortions that remain within the model are the impact of stakes on actions, non-random selection of participants into the experiment, and context effects. These three considerations are addressed in turn, with much of the discussion centered around individual bidding behavior in auctions since this area represents a vast and expanding literature.[28]

Stakes and cognitive costs

If there is a cost associated with effort, individuals' behavior should more closely match the predictions of rational-behavior theories as (1) the stakes of the decision increase, and (2) the decision costs decrease.[29] Smith and Walker (1993) find evidence of these two effects in a comprehensive review of thirty-one published laboratory experiments. Camerer and Hogarth (1999) extend Smith and Walker's survey by examining 74 experimental papers, and find evidence in favor of the cognitive-effort theory, noting that "higher levels of incentives have the largest effects in judgment and decision tasks." List and Lucking-Reiley (2000) explore these

[28] For example, in common value auctions the winner's curse has been a much discussed finding (see Thaler, 1992). The Winner's Curse (WC) represents a disequilibrium behavior in which bidders systematically overbid and thus earn a negative payoff upon winning. The phenomenon arises because bidders fail to take into account the fact that *if they win* then they may have over-estimated the value of the object, and correct their bids for that fact. A first illustration of the phenomenon is due to Bazerman and Samuelson (1983), who used a classroom jar experiment whereby an auction was conducted for a jar of coins worth $8.00, with prizes for the best guess. As illustrated in the survey of Kagel and Levin (2002), subsequent laboratory experiments have refined the analysis and have shown that the WC is "alive and well" across student subject pools including super-experienced" bidders (Kagel and Levin, 1999) and is a robust phenomenon in many common value auction forms (see, e.g., Kagel et al., 1995; Levin et al., 1996). Charness and Levin (2005, pp. 4-5) note, "It seems that presently only a few economists doubt the existence of WC-type of behavior in the lab and more and more are willing to accept that such behavior may indeed exists outside the lab in real markets."

[29] Note that this stands in contrast to games in which morality and wealth maximization are at odds with one another.

issues with auction experiments in the field, finding results that indicate that stakes matter—high-priced auctioned goods produced more of the theoretically predicted strategic behavior than did lower-priced goods.

Non-random selection of participants

If the participants in laboratory studies differ in systematic ways from the actors engaged in the targeted real-world settings, lab results can be misleading. This concern is particularly acute in settings related to environments where (1) the real-world stakes are both high, and (2) it is frequently groups of individuals, or firms, rather than individuals making such choices. In these cases, actual decision makers are likely to be a highly select group of individuals with expertise and experience. Lab participants lacking such experience and expertise might not provide accurate guidance as to how real decision-making occurs.

One piece of laboratory evidence that suggests selection effects of this sort may be important comes from Cox et al. (2001), who use a laboratory experiment with free entry and exit to show that only a subset of potential bidders elected to bid in auctions with a Winner's Curse (WC) element possible. The worst offenders chose to stay out, largely mitigating the WC problem.[30] Relatedly, if real market experience is important, then student behavior might not provide an adequate ideal for extrapolation purposes. For instance, List and Lucking-Reiley (2000) report that more experienced auction bidders exhibit a greater tendency to behave according to strategic theories than did lesser experienced bidders.[31]

[30] Selection effects have been shown elsewhere as well. In a study measuring individual risk propensities, Harrison et al. (2005) find that the use of show-up fees to increase experimental participation generates a sample of more risk averse participants.

[31] A standard counter-argument made by experimentalists to criticisms such as the one made here about novices versus experts is that allowing subjects to engage in multiple rounds of the same game or activity provides them with the chance to adapt to the environment. While for some tasks laboratory learning might be adequate, the data in List (2003) suggest that useful cognitive capital builds up slowly, over days or years, rather than in the short-run of an experiment. This observation implies that certain multiple-trial laboratory experiments may not provide economic agents with the necessary experience to overcome short-run anomalies.

The importance of context

Subjects, particularly experienced agents, come to experiments with rules of behavior learned in the outside world. Depending on whether the specific context of the lab rewards or punishes these rules of thumb, radically divergent results can be obtained. Harrison and List (2005), for instance, examine the behavior of professional bidders in their naturally occurring environments. In their real-world bidding, they do not fall prey to the winner's curse. When the expert bidders are placed in unfamiliar roles, however, they often fall prey to the winner's curse, just as happens in the lab.[32] This finding underscores that agents with decisionmaking defects will tend to disappear from markets and highlights that if the lessons learned outside of the lab fail miserably in the exogeneously imposed role, then professionals will likely perform worse than students.[33] This result is one of the important insights in Burns (1985), who reports that students do better than businessmen in progressive oral auctions. This result occurs because the wool traders used their related experience in the actual wool markets with which they operate, even in cases where such heuristics were inappropriate. Burns (1985, p. 150) summarizes this

[32] Framed field experiments are a step beyond artefactual field experiments and toward naturally-occurring situations in that they use a non-standard subject pool that has experience in the task, environment, stakes, and/or information set, commodity, etc. (see http://www.arec.umd.edu/fieldexperiments/ for a bibliography of more than 50 framed field experiments). Our intuition is that demand effects become less important as one increases the total environmental contingencies faced by the experimental subjects. Under this view, such effects become less important as one moves from a traditional laboratory experiment to a framed field experiment. This intuition is consonant with the model of Rosnow and Aiken (1973).

[33] Another means for the WC to be attenuated in naturally-occurring markets is for the market to develop securities to protect bidders from such defects. One example of this can be found in Dyer and Kagel (1996), who review how executives in the commercial construction industry appear to avoid the winner's curse in the field (p. 1464): "Two broad conclusions are reached. One is that the executives have learned a set of situation-specific rules of thumb which help them to avoid the winner's curse in the field, but which could not be applied in the laboratory markets. The second is that the bidding environment created in the laboratory and the theory underlying it are not fully representative of the field environment. Rather, the latter has developed escape mechanisms for avoiding the winner's curse that are mutually beneficial to both buyers and sellers and which have not been incorporated into the standard one-shot auction theory literature." This passage highlights that creating an artificial environment and executing a standard laboratory common value auction using the executive experts as subjects will likely lead to students outperforming the experts because the situation-specific rules of thumb that the experts have developed will not help them avoid the WC. Furthermore, even if the WC was rampant in this environment, it would not influence market outcomes because the market has endogenously created "escape mechanisms" to allow bidders to avoid repercussions of the WC.

insight as follows: "In general, where the rules of the experimental market conflicted with the market rules that they were used to, the buyers instinctively chose the latter."[34]

More generally, the importance of context in this class of games is illustrated in the vast research on the Wason selection task (Hertwig and Ortmann, 2001). The Wason selection task is often used to assess what information is necessary in order to test the truth of an abstract logical reasoning problem: if P then Q. A typical experiment will present a rule and ask subjects whether the rule is being violated. Consider the rule: *If a card has a J on one side, then it has a 5 on the other side.* Subjects are aware that on the particular set of cards, each one has a letter on one side and a number on the other side. Four cards are shown, a P, P', Q, and Q' card and subjects are asked to pick correctly the two cards to turn over to verify the rule. Roughly 1 in 5 subjects answer correctly: to turn over the P and Q' cards. When an identical structured problem is given a context: *If Derek hits the golf ball, then it will go in the water,* the success rate skyrockets as most subjects readily realize that they need to know what happened when Derek hit the ball (P) and the action associated with the ball not going in the water (Q').

IV. Implications of our model for experiments that measure group differences

In recent years, a series of laboratory experiments have uncovered systematic differences in behavior across broad groups of actors (e.g., by gender, race, or caste), as well as differential responses of members of these groups to the presence or absence of members of other groups. Croson and Gneezy (2005) review the former lot of studies in economics and report that women are generally more risk averse than men. In addition, they report that men systematically select into more competitive environments, even after conditioning on ability, than women. In a

[34] More generally, students tend to outperform experts in more cognitively challenging and abstract tasks. This result is illustrated in Cooper et al. (1999), who argue that students appear to outperform experts because they are more likely to have the "test-taking" skills such as accurately computing cognitively challenging calculations, following abstract chains of logic, and the like, that is required in their laboratory game.

related spirit, Steele (1997) reports that "stereotype threat" induces minority group members to perform poorly if the stereotype is cued by the experimenter. For instance, Spencer et al. (1999) report that activating the negative stereotype that female students are inferior to male students in math-related tasks leads to males outperforming females on math tests; whereas no such difference was observed when the stereotype was not activated. Likewise, Steele and Aronson (1995) examined these processes among African American students and found similar results when Black and White students took a verbal GRE exam.[35]

A second set of studies highlights the importance of group composition (and comparison) in performance. In Katz et al. (1964), black participants performed better on an IQ test when they believed their results would be compared to other Blacks, as opposed to other Whites. In Hoff and Pandey's (2004) study on members of low castes in India, in one treatment, prior to completing a task, the subjects are each asked to state the caste to which they belong. There were no caste differences in performance when caste was not publicly revealed, but making caste salient created a large caste gap. These results are generally interpreted as evidence that, for a variety of possible reasons, members of low status or marginalized groups have been conditioned to perceive that the "right" thing to do when paired with members of the dominant group is to perform poorly.

Our model, however, suggests a number of potential channels through which these two classes of laboratory findings may yield a distorted or exaggerated view of the importance of such behavior outside the lab. First, as we have stressed throughout this paper, the act of being watched and the emphasis on process are likely to greatly increase the weight that subjects place on the "right" behavior relative to the wealth-maximizing behavior. If the design of the

[35] Research due to Irwin Katz and his colleagues (e.g., Katz, 1964) can also be interpreted as manipulation of stereotype threat.

experiment signals to the subjects that the experimenter desires that minority groups underperform the dominant group, members of both groups may respond. Note that such experimenter demand effects lead to a different set of predictions than models such as stereotype threat. Under stereotype threat, cues suggesting that women or minorities do badly on the test should lead women or minorities to perform poorly, but should not affect the performance of men or whites. In contrast, if the operative force is that subjects are trying to please the experimenter, then cues that men/whites are better at a test than women/minorities will cause women and minorities to do poorly *and* will lead men and whites to perform better. Indeed, in Spencer et al.'s (1999) male/female math test study, before subjects took the exam, in one treatment the experimenter informed the subjects that the test generally showed gender differences, and the women tested worse than men. In another treatment, the experimenter informed the subjects that the test generally showed no gender differences. Consistent with our argument that the subjects were attempting to please the experimenter, the convergence in test scores occurred not only because female's performance increased, but also because while male's performance *decreased*. (i.e., males performed better when they were supposed to outperform females and they performed worse when they were supposed to be isomorphic to females—see Figure 1 in Steele (1997)). Similar data patterns are observed in the race study experiments of Steele and Aronson (1995—see Figures 2 and 3 in Steele (1997) and Figures 2 and 4 of Steele and Aronson (1995)).

Another important factor that our model stresses is the potential systematic differences across groups in the type of subjects who self-select into experiments. If, for instance, certain subjects are especially likely to participate in experiments, the results obtained may not be representative of group differences as a whole if the selection process is not orthogonal to group

type. Consider the results presented on risk aversion across men and women. Many lab studies report that women are more risk averse than men (see, e.g., Holt and Laury (2002) and the Croson and Gneezy (2005) survey). Yet, two recent studies highlight some evidence that is suggestive that such differences might be exaggerated.

Harrison et al. (2005) explore risk propensities amongst a representative sample of the Danish population. Most importantly for our purposes, their study yielded an extremely high participation rate: 96% of those invited to the experiment agreed. Using identical protocol as those used in studies in the U.S. that report gender differences (e.g., Holt and Laury, 2002), Harrison et al. (2005) find no gender differences in their sample. Of course, this could be due to several reasons, including spatial differences in risk posture, sample composition, statistical techniques employed (as argued in Harrison et al. (2005)), or selection effects. Consonant with the latter interpretation, Dwyer et al. (2003) use data from a national survey of almost 2,000 mutual fund investors to investigate whether investor gender is related to risk taking as revealed in mutual fund investment decisions. They report that women are more risk averse than men, but they find that the impact of gender on revealed risk preference is considerably weakened when investor knowledge is controlled in the regression equation.[36]

Another recent example of the importance of selection effects is highlighted in Casari et al. (2005). Comparing bidding of various types of agents in common value auctions, the authors (p. 1) report that "there are strong selection effects in bid estimates for both experienced and inexperienced subjects…..ignoring these selection effects is most misleading for inexperienced

[36] A related point on selection that pervades not only this literature but the experimental literature more broadly is that the lab might yield a biased treatment effect if there is a treatment-demographic interaction and the lab misses the important demographic. Anderson et al. (2005) illustrate this point by considering preferences over risk and time. They find that there (p. 1) "are also differences in treatment effects measured in the lab and the field that can be traced to interactions between treatment and demographic effects. These can only be detected and controlled for properly in the field data."

bidders, as the unbiased estimates of the bid function indicate much faster learning and adjustment…than do the biased estimates." This result underscores that differences across bidder type are exacerbated if the selection effect is not accounted for in the analysis. Similar selection effects are found in Rutström (1998), who explored the role of recruitment fees in securing experimental participants and found some evidence of group differences (male versus female, white versus non-white) in bidding in auctions in some of the treatments and no differences in other treatments, consistent with our model.

V. Concluding Remarks

In this paper, we provide an analysis of when findings from the lab are, or are not, likely to provide reliable inferences outside of the laboratory. We argue that games in which the moral and wealth-maximizing choices are in conflict (e.g., experiments measuring social preferences), are likely to yield experimental results which diverge most sharply from real-world behavior. The empirical evidence suggests that being watched, along with the emphasis on process, systematically distorts behavior in the laboratory in the direction of placing more weight on the moral action rather than the wealth-maximizing one. In games where morality is not an issue, low stakes, non-random selection of participants (both in the lab and among those who perform a task in real life), and details of whether a game rewards or punishes the heuristics used by experts all influence the experimental findings. Differences observed across groups of race, sex, and age may similarly be exaggerated by the laboratory environment, although less empirical evidence is currently available on this question.

Based on our analysis, we envision an important, but more limited, role for traditional laboratory experiments in economics. Lab experiments can provide a crucial first understanding of treatment effects that *can* happen, and they usefully provide an underlying mechanism that

might be at work when certain data patterns are observed. In this light, experimenters are like aerodynamicists who use wind tunnels to test models of proposed aircraft, helicopters, cars, and trains. The wind tunnel provides the engineer with valuable data on scale models much like the lab provides economists with important insights on an economic phenomenon.[37]

We view the experimental literature as usefully extending in several directions at this point, of which we discuss three. First, combining laboratory analysis with a model of decision-making, such as the simple one we present in this paper, expands the potential role of lab experiments. By anticipating the types of biases common to the lab, experiments can be designed to minimize such biases (e.g. through attempting to ensure subject anonymity). Further, knowing the sign and plausible magnitude of any biases induced by the lab, one can extract useful information from a study, even if the results cannot be seamlessly extrapolated outside the lab. Second, by focusing on qualitative rather than quantitative insights much can be learned.[38] In games that pit wealth and morality, for example, one avenue is to "nest" treatments and perform a difference-in-difference estimation that effectively "nets" out laboratory effects. Finally, before the dust settles on several of the issues that we raise—which we argue remain open empirical questions—a bridge must be created that links laboratory experiments and reality. We have confidence that field experiments can provide the necessary mix of control and realism that can put many of these issues to rest.

[37] Some argue that the wind tunnel is the most lasting contribution of the Wright brothers to the science of aerodynamics. Interestingly, it is estimated that it took the Wright Brothers less than 20 hours of wind tunnel testing to produce their successful flyer.

[38] For instance, Levine and Plott (1977) report that agenda influences are directionally consistent across laboratory experiments and a naturally-occurring setting that they manipulated to secure a more favorable outcome in their flying club.

References

Akerlof, George A., "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics,* November 1982, *97*(4), pp. 543-569.

Akerlof, George A. and Rachel E. Kranton. "Economics and Identity," *Quarterly Journal of Economics*, August 2000, *115*(3), pp. 715-753.

Akerlof, George A and Rachel E. Kranton. "Identity and the Economics of Organizations." J*ournal of Economic Perspectives*, Winter 2005, *19*(1), pp. 9-32.

Anderson, Steffen, Glenn W. Harrison, Morten Igel Lau and E. Elisabet Rutström, "Preference Heterogeneity in Experiments: Comparing the Field and Lab," working paper, University of Central Florida, 2005.

Andreoni, James. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy*, December 1989, *97*, pp. 1447-58.

Andreoni, James. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal*, June 1990, *100*, pp. 464-477.

Andreoni, James and Miller, John, "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism, " *Econometrica*, March 2002, *70*(2), pp. 737-753.

Andreoni, James, and Vesterlund, Lise, "Which Is the Fair Sex? Gender Differences in Altruism," *Quarterly Journal of Economics*, February 2001, *116*(1), pp. 293-312.

Ashraf, Nava, Camerer, Colin F., and Loewenstein, George. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives*, Summer 2005, *19*, no. 3, forthcoming.

Bandiera, Oriana, Iwan Rasul, and Imran Barankay. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." Forthcoming *Quarterly Journal of Economics.*

Bardsley, Nicholas. "Altruism or Artefact? A Note on Dictator Game Giving," University of Nottingham working paper.

Bazerman, M.H. and W.F. Samuelson. "I Won the Auction But Don't Want the Prize." *Journal of Conflict Resolution*, 1983, *27,* pp. 618-634.

Becker, Gary S. "A Theory of Social Interactions," *Journal of Political Economy*, 1974, *82*(6), pp. 1063-93.

Benz, Matthias and Stephan Meier. "Do People Behave in Experiments as in real Life? Evidence from Donations," University of Zurich, working paper, 2005.

Berg, Joyce; Dickhaut, John W. and McCabe, Kevin A., "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, July 1995, *10*(1), pp. 122-142.

Bernheim, B.D. "A Theory of Conformity." *Journal of Political Economy*, 1994, *102*(5), pp. 841-877.

Binmore, Ken, *Game Theory and the Social Contract. Volume1. Playing Fair,* Cambridge and London: MIT Press, 1994.

Binzel, Richard P., "Pluto." *Scientific American*, 1990, *262*, pp. 50-56.

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

Brown, Martin, Armin Falk, and Ernst Fehr. "Relational Contracts and the Nature of Market Interactions," *Econometrica*, 2004, *72*(3), pp. 747-780.

Brunswik, Egon. "Representative Design and Probabilistic Theory in a Functional Psychology," *Psychological Review*, 1955, 62(3), pp. 193-217.

Brunswik, Egon. *Perception and the Representative Design of Psychological Experiments* (2nd ed.). Berkeley, CA: University of California Press, 1956.

Burns, Penny, "Experience and Decisionmaking: A Comparison of Students and Businessmen in a Simulated Progressive Auction," in *Research in Experimental Economics*, (JAI), ed. Vernon L. Smith, pp. 139-153 (1985).

Camerer, Colin F., "Behavioral Game Theory: Experiments in Strategic Interaction, " 2003, pp. xv, 550, *Roundtable Series in Behavioral Economics*. Princeton: Princeton University Press

Camerer, Colin F. and Hogarth, Robin M., "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 1999, (19), pp. 7-42.

Camerer, Colin F. and Weigelt, Keith, "Experimental Tests of a Sequential Equilibrium Reputation Model," *Econometrica*, January 1988, *56*(1), pp. 1-36.

Cameron, Lisa A. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Economic Inquiry*, January 1999, *37*(1), pp. 7-59

Campbell, D. T., & Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin, 1963.

Carpenter, Jeffrey, Verhoogen, Eric, and Burks, Stephen. "Comparing Students to Workers: The Effects of Stakes, Social Framing, and Demographics on Bargaining Outcomes," in *Field*

Experiments in Economics, (JAI), eds. Jeffrey Carpenter, Glenn Harrison, John A. List, pp. 261-290 (2005).

Carpenter, Jeffrey, Verhoogen, Eric, and Burks, Stephen. "The Effect of Stakes in Distribution Experiments," *Economics Letters*, March 2005, *86*(3), pp. 393-98.

Casari, Marco, Ham John C., and Kagel, John H. "Selection Bias, Demographic Effects and Ability Effects in Common Value Auction Experiments," Ohio State University working paper, 2005.

Charness, Gary, "Attribution and Reciprocity in an Experimental Labor Market," mimeo, 1996.

Charness, Gary and Levin, Dan, "The Origin of the Winner's Curse: A Laboratory Study," 2005, UCLA Department of Economics.

Charness, Gary and Rabin, Matthew, "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), August 2002, 817-869.

Collard, David. *Altruism and Economy: A Study in Nonselfish Economics*. New York, NY: Oxford University Press, 1978.

Cooper, David J., "Gaming against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers." *American Economic Review*, September 1999, *89*(4), pp. 781-804.

Cox, James C.; Dinkin, Sam and Swarthout, James T., "Endogenous Entry and Exit in Common Value Auctions," *Experimental Economics*, October 2001, *4*(2), pp. 163-81.

Croson, Rachel and Uri Gneezy. "Gender Differences in Preferences." *Mimeo*, Wharton School, University of Pennsylvania, 2005.

Cross, J. "Some Comments on the Papers by Kagel and Battalio and by Smith," in J. Kmenta and J. Ramsey (eds.) Evaluation of Econometric Models, New York: New York University Press, 1980.

Cummings, Ronald G. and Laura O. Taylor. "Does Realism Matter in Contingent Valuation Surveys?" *Land Economics* 1998*,* 74: 203-15.

Diamond, Peter and Hausman, Jerry. "Contingent Valuation: Is Some Number Better than No Number?" *Journal of Economic Perspectives*, 1994, 8, pp. 45-64.

Dwyer, Peggy D., James H. Gilkeson, and John A. List. "Gender Differences in Revealed Risk Taking: Evidence From Mutual Fund Investors." *Economics Letters,* 2002, *76*(2), pp. 151-158.

Eckel, Catherine and Phillip Grossman. "Altruism in Anonymous Dictator Games," *Games and Economic Behavior,* 1996, *16*, pp. 181-191.

Eckel, Catherine C. and Wilson, Rick K., "Is Trust a Risky Decision?" *Journal of Economic Behavior and Organization*, December 2004, *55*(4), pp. 447-65.

Fehr, Ernst and Armin Falk, Wage Rigidity in a Competitive Incomplete Contract Market, *Journal of Political Economy*, 1999 (107), pp. 106-134.

Fehr, Ernst and Simon Gächter, "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 2000, 14(3), pp. 159-181.

Fehr, Ernst; Gachter, Simon and Kirchsteiger, Georg, "Reciprocity as a Contact Enforcement Device: Experimental Evidence," *Econometrica*, July 1997, 65(4), pp. 833-60.

Fehr, Ernst, Kirchsteiger, George; Riedl, Arno; "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, May 1993, *108*(2), pp. 437-59.

Fehr, Ernst and List, John A., "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOS, " *Journal of the European Economic Association*, September 2004, *2*(5), pp. 743-71.

Fehr, Ernst and Schmidt, Klaus, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 1999, *114*, pp. 817-868.

Fehr, Ernst, Urs Fischbacher, and E. Tougareva, "Do High Stakes and Competition undermine Fairness - Evidence from Russia," Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 120.

Feynman, R.P., R.B. Leighton and M. Sands *The Feynman Lectures on Physics*. Reading, MA: Addison-Wesley Publishing Company, 1963.

Forsythe, Robert, et al., "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior*, May 1994, *6*(3), pp. 347-69.

Frey, Bruno S., Felix Oberholzer-Gee, and Reiner Eichenberger, "The Old Lady Visits Your Backyard: A Tale of Morals and Markets." *The Journal of Political Economy*, 1996, *104*(6), pp. 1297-1313.

Gachter, Simon and Falk, Armin, "Reputation and Reciprocity: Consequences for the Labour Relation," *Scandinavian Journal of Economics*, 2002, *104*(1), pp. 1-26.

Gintis, Herbert, "The Contribution of Game Theory to Experimental Design in the Behavioral Sciences," *Behavioral and Brain Sciences*, 2001, 24:411-412.

Grampp, William. 1948. "Adam Smith and the Economic Man." *The Journal of Political Economy*. August 1948, *56*:4, pp. 315-336.

Hanemann, W. Michael, "Willingness to Pay and Willingness to Accept: How Much Can They Differ?" *American Economic Review*, June 1991, *81*(3), pp. 635-47.

Haney, C., Banks, W. C., Zimbardo, P. G. "Interpersonal Dynamics in a Simulated Prison." *International Journal of Criminology and Penology* 1973, *1*, pp. 69–97.

Hannan, R. Lynn, John H. Kagel, and Donald V. Moser. "Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior," *Journal of Labor Economics,* 2004, *20*(4), pp. 923-951.

Harbaugh, William T. and Krause, Kate, and Steven Liday, "Bargaining by Children," University of Oregon working paper, 2003.

Harbaugh, William T., Kate Krause, Steven G. Liday, and Lise Vesterlund. "Trust in Children." In Elinor Ostrom and James Walker eds., *Trust, Reciprocity and Gains from Association: Interdisciplinary Lessons from Experimental Research*. New York City, NY: Russell Sage Foundation, 2003.

Harrison, Glenn, Morten Lau, and Elisabet Rutstrom. "Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments." UCF working paper, 2005.

Harrison, Glenn W. and List, John A., "Field Experiments," *Journal of Economic Literature*, December 2004, *42*(4), pp. 1009-55.

Harrison, Glenn W. and John A. List. "Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of Winner's Curse," UCF working paper, 2005.

Harrison, Glenn and Elisabet Rutstrom. "Doing It Both Ways -- Experimental Practice and Heuristic Context" *Behavioral and Brain Sciences*, 2001, *24*(3), pp. 413-414.

Heiner, R. "The Origin of Predictable Behavior," American Economic Review, 1983, *73 (4),* pp. 560-595.

Henrich, J. et al. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-scale Societies." *American Economic Review*, May 2001, *91* (2), pp. 73-78.

Henrich, J., R. Boyd, S. Bowles, H. Gintis, E. Fehr, C. Camerer, R. McElreath, M. Gurven, K. Hill, A. Barr, J. Ensminger, D. Tracer, F. Marlow, J. Patton, M. Alvard, F. Gil-White, and N. Smith. "'Economic Man' in Cross-Cultural Perspective: Ethnography and Experiments from 15 small-scale societies." Forthcoming in *Behavioral and Brain Sciences.*

Hertwig, Ralph and Andreas Ortmann, "Experimental Practices in Economics: A Methodological Challenge for Psychologists," *Behavioral and Brain Sciences*, 2001, 24:383-451.

Hoff, Karla, and Priyanka Pandey, "Belief Systems and Durable Inequalities: An Experimental Exploration of Indian Caste," World Bank Working Paper No. 3351, 2004.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 1994, *7*(3), pp 346-380.

Hoffman, Elizabeth; McCabe, Kevin and Smith, Vernon L.; "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review*, June 1996, *86*(3), pp. 653-60

Holt, Charles A. "Markets, Games, and Strategic Behavior: Recipes for Interactive Learning," 2005, Addison-Wesley, forthcoming.

Holt, Charles A. and Susan K. Laury. "Risk Aversion and Incentive Effects." *American Economic Review*, 2002, *92*(5), pp. 1644-1655.

Jones, Edvard E. *Social stigma – The Psychology of Marked Relationships*. New York, NY: W. H. Freeman, 1984.

Jones, Stephen. *The Economics of Conformism*. Oxford & New York: Basil Blackwell, 1984.

Kagel, John H., Raymond C. Battalio, and James M. Walker, "Volunteer Artifacts in Experiments in Economics; Specification of the Problem and Some Initial Data from a Small-Scale Field Experiment," in *Research in Experimental Economics*, (JAI), ed. Vernon L. Smith, pp. 169-197 (1979).

Kagel, J.H., Levin, D., and Harstad, R. "Comparative Static Effects of Number of Bidders and Public Information on Behavior in Second-Price Common-Value Auctions." *International Journal of Game Theory*, 1995, *24*, pp. 297-319.

Kagel, John H. and Levin, Dan, "Common Value Auctions with Insider Information," *Econometrica*, September 1999, *67*(5), pp. 1219-38

Kagel, John H. and Levin, Dan, "Common Value Auctions and the Winner's Curse," 2002, pp. xvi, 401, Princeton and Oxford: Princeton University Press

Kahneman, Daniel, and Tversky, Amos. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47 (March 1979): 263-91.

Karni, Edi. "Impartiality: Definition and Representation." *Econometrica* 1998, *66*, pp. 1405–1415.

Katz, Irwin. "Review of Evidence Relating to Effects of Desegregation on the Intellectual Performance of Negroes." *American Psychologist*, 1964, *19* (6), pp. 381-399.

Katz, Irwin, E.G. Epps, and L.J. Axelson. "Effect upon Negro Digit-Symbol Performance of Anticipated Comparison with Whites and with Other Negroes." *Journal of Abnormal and Social Psychology*, 1964, *69,* pp. 963-70.

Kotchen, Matthew J. "Impure Public Goods and the Comparative Statics of Environmentally Friendly Consumption." *Journal of Environmental Economics and Management*, March 2005, *49*(2), pp. 281-300.

Laffont, Jean-Jacques. "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics," *Economica*, 1975, *42*(168), pp. 430-437.

Ledyard, J. O. "Public Goods: A Survey of Experimental Research." In ed. by J. Kagel and A. Roth *Handbook of Experimental Economics*, NJ: Princeton University Press, 1995.

Levin, Dan, John H. Kagel, and Jean-Francois Richard. "Revenue Effects and Information Processing in English Common Value Auctions." *The American Economic Review*, 1996, *86*(3), pp. 442-460.

Levine, Michael E. and Charles R. Plott. "Agenda Influence and its Implications," *Virginia Law Review*, (1977), 63, pp. 561-604.

List, John A. "Does Market Experience Eliminate Market Anomalies?," *Quarterly Journal of Economics* (2003), 118(1), pp. 41-71.

List, John A. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, forthcoming, 2005.

List, John A. and Todd L. Cherry. "Examining the Role of fairness in High Stakes Allocation Decisions," *Journal of Economic Behavior and Organization*, 2005, forthcoming.

List, John A. and Lucking-Reiley, David. "Demand Reduction in a Multi-Unit Auction: Evidence from a Sportscard Field Experiment," *American Economic Review*, 2000, September, 90(4), pp. 961-972.

List, John A., Robert Berrens, Alok Bohara, and Joe Kerkvliet. "Examining the Role of Social Isolation on Stated Preferences." *American Economic Review*, 2004, *94*(3), pp. 741-752.

Loewenstein, George, "Experimental Economics from the Vantage-Point of Behavioural Economics," *Economic Journal*, February 1999, *109*(453), pp. F23-34

Loomes, Graham and Sugden, Robert, "Regret Theory: An Alternate Theory of Rational Choice under Uncertainty," *Economic Journal*, December 1982, *92*(368), pp. 805-24.

Margolis, Howard. *Selfishness, Altruism, and Rationality.* Chicago, IL: University of Chicago Press, 1982.

Masclet, David et al. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism," American Economic Review, 2003, *93*(1), pp. 366-380.

Matthias Benz and Stephan Meier, 2005. "Do People Behave in Experiments as in Real Life? – Evidence from Donations," Institute for Empirical Research in Economics - Working Papers iewwp248, 2005.

Milgram, Stanley. "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology*, 1963, *67*, pp. 371-378.

Mook, D. G. "In Defense of External Invalidity." *American Psychologist,* 1983, *38*, pp. 379-387.

Mullainathan, Sendhil and Thaler, Richard H., "Behavioral Economics, " *National Bureau of Economic Research,* Inc, 2000, NBER Working Papers: 7948

Newton, Isaac. *Mathematical Principles of Natural Philosophy*. Los Angeles, CA: University of California Press, 1966.

Orne, M. T. *The Demand Characteristics of an Experimental Design and their Implications*. Paper read at American Psychological Association, Cincinnati, 1959a.

Orne, M. T. "The Nature of Hypnosis: Artifact and Essence." *Journal of Abnormal and Social Psychology,* 1959b, *58,* pp. 277-299.

Orne, Martin T. "On the Social Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist*, *17*(10), 1962, 776-783.

Ortmann, Andreas and Ralph Hertwig. "Why Anomalies Cluster in Experimental Tests of One-shot and/or Finitely Repeated Games: Some Evidence from Psychology." working paper, 2000.

Parco, J. E., Rapoport, A., & Stein, W. E. "Effects of Financial Incentives on the Breakdown of Mutual Trust." *Psychological Science*, 2002, *13*, pp. 292-297.

Pierce, A. H. "The Subconscious Again." *Journal of Philosophy, Psychology, & Scientific Methods,* 1908, *5*, pp. 264-271.

Plott, Charles R. "Industrial Organization Theory and Experimental Economics." *Journal of Economic Literature,* 1982, *20*, pp. 1485-1527.

Potters, Jan and van Winden, Frans, "Professionals and Students in a Lobbying Experiment: Professional Rules of Conduct and Subject Surrogacy," *Journal of Economic Behavior and Organization*, December 2000, *43*(4), pp. 449-552.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, December 1993, *83*(5), pp. 1281-1302.

Rabin, Matthew, "Psychology and Economics," *Journal of Economic Literature*, March 1998, *36*(1), pp. 11-46.

Rege, Mari and Kjetil Telle. "The impact of social approval and framing on cooperation in public good situations," *Journal of Public Economics*, 2004, *88*(7-8), pp. 1625-1644.

Rosenthal, R.L. "Covert Communication in the Psychological Experiment." *Psychological Bulletin,* 1967, *67,* pp. 356–367.

Rosenthal, R.L. and R.L. Rosnow, *Artifact in Behavioral Research*, New York: Academic Press, 1969.

Rosenthal, R.L. and R.L. Rosnow, *The Volunteer Subject*, New York: John Wiley and Sons, 1973.

Rosenthal, Robert W., "Market Games with Production and Public Commodities," 1976, Northwestern University, Center for Mathematical Studies in Economics and Management Science, Discussion Papers: 156

Rosenthal, Robert W., "A Note on Robustness of Equilibria with Respect to Commitment Opportunities," *Games and Economic Behavior*, May 1991, *3*(2), pp. 237-43.

Rosnow, R.L. and L.S. Aiken, "Mediation of Artifacts in Behavioral Research," *Journal of Experimental Social Psychology*, 1973, 9: pp. 181-201.

Ross, L. and Ward, A. "Naive Realism: Implications for Social Conflict and Misunderstanding." in T. Brown, E. Reed, and E. Turiel eds., *Values and Knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1996, pp. 103-135.

Roth, Alvin E. "Bargaining Experiments." in Kagel, J.H., Alvin E. R., eds., *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press, 1995, pp. 253–342.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *The American Economic Review*, 1991, *81*(5), pp. 1068-1095.

Rutstrom, Elisabet. "Home-grown Values and Incentive Compatible Auction Design", *International Journal of Game Theory*, 1998, *27*(3), pp. 427-441.

Samuelson, Paul. Discussion Contribution. *American Economic Review*, 1963, *53*(2), pp. 231-236.

Schmuckler, M. A. "What is Ecological Validity? A Dimensional Analysis." *Infancy*, 2001, *2,* pp. 419-436.

Schram, Arthur. "Artificiality: The Tension Between Internal and External Validity in Economic Experiments." *Journal of Economic Methodology*, 2005, *12(2)*, pp. 225-237.

Sen, Amartya K. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Journal of Philosophy and Public Affairs*, 1977, *6*, pp. 317-44.

Settle, Thomas B "An Experiment in the History of Science". *Science*, 1961, *133*, pp. 19-23.

Shapley, Harlow. *Of Stars and Men*: *Human Response to an Expanding Universe*. Westport CT: Greenwood Press, 1964.

Slonim, Robert and Roth, Alvin E., "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic, " *Econometrica*, May 1998, *66*(3), pp. 569-96

Smith, Adam. *The Theory of Moral Sentiments.* eds. D.D. Raphael and A.L. Macfie. Indianapolis: Liberty Fund, 1982.

Smith, Vernon L. "Microeconomic Systems as an Experimental Science." *The American Economic Review*, 1982, *72*(5), pp. 923-955.

Smith, Vernon L. and Walker, James M., "Rewards, Experience and Decision Costs in First Price Auctions," *Economic Inquiry*, April 1993, *31*(2), pp. 237-45

Sobel, Joel (2002). "Social Preferences and Reciprocity," mimeo, U.C. San Diego.

Spencer, S. J., C.M. Steele, and D.M. Quinn. "Stereotype Threat and Women's Math Performance." *Journal of Experimental Social Psychology*, 1999, *35*, pp. 4-28.

Starmer, Chris. "Experimental Economics: Hard Science or Wasteful Tinkering?" *Economic Journal*, 1999a, *109(453)*, pp. F5-15.

Starmer, Chris. "Experiments in Economics: Should we Trust the Dismal Scientists in White Coats?" *Journal of Economic Methodology*, 1999b, 6*(1)*, pp. 1-30.

Steele, C. M. "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." *American Psychologist*, 1997, *52*, pp. 613-629.

Steele, C.M., J. Aronson. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology,* 1995, *69*, pp. 797-811.

Sugden, Robert. "Reciprocity: The Supply of Public Goods Through Voluntary Contributions." *Economic Journal* 1984, *94,* pp. 772-787.

Thaler, Richard H., "The Winner's Curse: Paradoxes and Anomalies of Economic Life," 1992, pp. ix, 230, Princeton and Chichester, U.K: University Press

Tversky, Amos and Kahneman, Daniel, "Loss Aversion in Riskless Choice: A Reference-Dependent Model," *Quarterly Journal of Economics*, November 1991, *106*(4), pp. 1039-61

Wilde Louis, “On the Use of Laboratory Experiments in Economics,” In Joseph Pitt, ed., *The Philosophy of Economics*. Dordrecht: Reidel, 1981, pp. 137-143.