

Experimenting with Fish has some Advantages

John A List
University of Chicago and NBER

17 December 2007

As I sat through the morning session of the Implications of Behavioral Economics for Economic Policy I was struck by the interesting presentations as well as the insightful spectator remarks. One particularly astute gentleman sitting directly at my side made several penetrating comments. One comment related to how one should use experimental methods in economics. He noted that there are distinct advantages in conducting experiments with humans rather than fish, for example, since we can ask humans how they came to their choice and probe into their interpretations of the situation. Of course, this is a valid point. The biologist studying animal behavior has no idea what it feels like to be a goldfish when the water temperature suddenly changes from 74° Fahrenheit to 80° Fahrenheit, and the fish might find it difficult to relay that information. Likewise, the chemist has little idea what it feels like to be Uranium₂₃₉ when it turns into Neptunium. Experimental economists are potentially in a more fortunate situation. We are able to study the behavior of people, where we can experience the situations ourselves and ask our experimental subjects about their own experiences.

At the heart of the burgeoning literature studying fairness in labor markets is experimental evidence of just such kind. The literature has produced an impressive array of experimental treatments that provide data that 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 rewards (see, e.g., Fehr et al., 1993; Berg et al., 1995; Charness, 1996; Fehr et al., 1997; Fehr and Falk, 1999; Fehr and Gächter, 2000; Gächter and

Falk, 2002; Hannan et al., 2004; Brown et al., 2004). Such findings have been argued to show that (Fehr and Gächter, 2000) “reciprocity has powerful implications for many important economic domains...There are important conditions in which self-interest theory is unambiguously refuted.....markets with incomplete contracts, the reciprocal types dominate the aggregate results.”

The study for which I have been asked to comment prescribes how labor market models should be changed to reflect the manner in which fairness perceptions might affect the labor market. The work uses laboratory experiments as the cornerstone of the scientific evidence. This is an important step in the discovery process, and I laud the authors for their progress. They have made important strides in this study, and in the broader literature with this research agenda. Yet, my assigned duty is not to heap praise on these scholars, but to discuss some issues at the heart of this most recent commotion.¹ In this regard, my comment will take a step back and consider more carefully the evidence—both from the lab and field—that has caused this ruckus in the economics community. I should stress at the beginning that I find this research agenda fascinating and that I firmly believe that certain agents have social preferences (see my own field work, for example, List and Lucking-Reiley, 2002, List, 2004, Landry et al., 2006; Karlan and List, 2007).

The main message of my comment is that the field evidence of social preferences from gift exchange games is more mixed than Fehr et al. (2007) conclude; and in fact when the data

¹ The interested reader should see the older literature as well. Kaufman (1988) provides an edited volume that reviews and assesses the work of four institutional labor economists (John Dunlap, Clark Kerr, Richard Lester, and Lloyd Reynolds). These four economists wrote a paper titled "Does The New Generation of Labor Economists Know More than the Old Generation?" for Richard Freeman in 1987. The piece provides a nice overview. Reese and Slichter are among other authors who might be of interest to scholars of this topic. One interpretation is that what separates the current interest in this topic from the older literature is due to the recent experimental evidence brought forth. Thanks to Alan Krueger for pointing me in this direction and providing the citations.

provide a clean measure of social preferences the effect is found to be small. Furthermore, interpretation from the field studies that do provide strong results in the gift exchange game is confounded—both reputational concerns and social preferences might be at work. Lab experiments potentially handle this confound, but slight perturbations of experimental conditions can dramatically alter behavior, and the important properties of the lab situation are not conducive to fluid generalizability to the extra-lab world. In this way, estimating deep preference parameters in the lab is lacking, especially in light of the fact that we have no theory to generalize such parameters.

I. A Framework

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. Recently, with Levitt (2007) I have argued that human decisions are influenced not just by monetary calculations, but also by at least six other factors: 1) The presence of moral and ethical considerations, 2) The nature *and* extent to which one's actions are scrutinized by others, 3) The particular context in which the decision is embedded, 4) Self-selection of the individuals making the decisions, and 5) The stakes of the game.

To make my basic point, I reconsider briefly the framework we introduced. 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, denoted v , the greater the decision's impact on W_i . The second effect is the non-pecuniary moral cost or benefit associated with action i , denoted as M_i . If, for instance, an individual has strong social

preferences, he will derive utility from charitable contributions.

In practice, many factors influence the moral costs associated with an action, but for modeling purposes, we focused on just three aspects of the moral determinant: i) the greater is the negative impact of an action on others, the more negative the moral payoff M_i , ii) the strength of social norms or legal rules that govern behavior in a particular society influence behavior, and iii) moral concerns depend on the nature and extent of how an individual's actions are scrutinized. Scrutiny is inherently multi-dimensional, but for simplicity assume that it includes only the nature and extent of scrutiny. An example of the nature of scrutiny is the presence of an experimenter, who potentially alters the subject's perception of the situation. More broadly, the experimental environment itself might draw upon a different set of expectations than markets. The extent of scrutiny relates to the anonymity of the subject's decision. In the model below, I denote the effect of scrutiny as s , with higher levels of s associated with greater moral costs.

Focusing on the case in which utility is additively separable in the moral and wealth arguments, I make the utility function when an individual i takes action a as

$$U_i(a, v, n, s) = M_i(a, v, n, s) + W_i(a, v) \quad (1)$$

Solving this simple decision problem yields several predictions, as discussed in Levitt and List (2007). For example, the greater is the social norm against the wealth maximizing choice or the degree of scrutiny, the larger the deviation from that choice. Further, as the stakes of the game rise, wealth concerns will increase in importance relative to fairness concerns, that is $|\partial M / \partial v| < |\partial W / \partial v|$. Such a framework makes it clear that the greater the extent that the lab environment mirrors the naturally occurring setting that it is modeling, the more confident one can be that the lab results will be generalizable. If the lab diverges from the environment of interest, the model provides a framework for predicting in what direction behavior in the lab will

deviate from that outside the lab.

II. Empirical Evidence

The model can speak to a wide range of experimental results, but its bite is likely to be greatest for those games in which there is the potential for a strong moral component to behavior. Research on social preferences, one of the most influential areas in experimental economics in recent years, fits this bill. Figure 1 highlights a handful of popular empirical approaches—ranging from methods that generate data to techniques used to model data—that have been used to explore preferences. In the westernmost portion of Figure 1 is laboratory experiments, which are used to generate data, and by their construction the *ideal* laboratory experimental environment represents the “cleanest test tubes” case. Some might view sterility as a necessary detraction, but sterility serves an important purpose: in an ideal laboratory experiment this very sterility allows an uncompromised glimpse at the effects of exogenous treatments on behavior in the lab. Of course, making generalizations outside of this domain might prove difficult in some cases, but to obtain the effect of treatment in this particular domain the only assumption necessary is appropriate randomization.

The easternmost part of the empirical spectrum in Figure 1 includes several examples of empirical models that make necessary identification assumptions to identify treatment effects from naturally-occurring data. These are well known and need not be further discussed here. Between laboratory experiments and models estimated using naturally-occurring data are the

various types of field experiments.² As discussed more fully in List (2006), field experiments represent a useful bridge between lab and naturally-occurring data.

A useful starting point to summarize the literature is to consider the findings using naturally-occurring data. An early excellent example is the striking evidence consistent with negative reciprocity on the part of disgruntled Firestone employees documented by Krueger and Mas (2004). Making use of variation in product quality induced by the contentious strike and subsequent hiring of replacement workers at Bridgestone/Firestone's Decatur, Illinois, plant in the mid-1990s, the authors report that labor strife at the Decatur plant closely coincided with lower product quality. Similarly, Mas (2006) documents persistent adverse effects on police performance following arbitration decisions in favor of the municipality.

The evidence using naturally-occurring data is not uniform, however. Chen (2005), who uses a large data set drawn from the Australian Workplace Industrial Relations Survey to explore reciprocity in the workplace, finds little evidence consistent with positive or negative reciprocity. In addition, empirical results in Lee and Rupp (2006) share similarities to the short run effects of shocks observed in Gneezy and List (2006). Lee and Rupp's (2006) examination of the effort responses of U.S. commercial airline pilots following recent pay cuts reveals that the behavior is very short-lived: even though in the first week after a pay cut frequent and longer flight delays are observed, after the first week there is no difference in airline flight performance.

² Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment. In doing so, they adopted the term "artefactual" field experiment to denote laboratory experiments with non-standard subject pools. Moving closer to how naturally-occurring data are generated, Harrison and List (2004) denote a "framed field experiment" as the same as an artefactual field experiment but with field context in the commodity, task, stakes, or information set of the subjects. Finally, a "natural field experiment" is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are participants in an experiment. Such an exercise is important in that it represents an approach that combines the most attractive elements of the lab and naturally-occurring data: randomization and realism.

Moving from the evidence drawn from naturally-occurring data to the results from field experiments that explore behavior in repeated play settings, I find the most recent evidence to be consistent with positive and negative reciprocity (see, e.g., List (2006), Bandiera et al. (2006), Al-Ubaydli et al. (2006), Bellemare and Shearer (2007), Cohn et al., (2007), and Marechal and Thoni (2007)). For instance, in List (2006) I had buying confederates approach dealers on the floor of a sportscard show, instructing them to offer different prices in return for sportscards of varying quality. When there was likely to be future interaction and the consumers could easily certify sportscard quality, I found a strong statistical relationship between price and quality provided. Similar data patterns are observed in the studies cited above as well as the experimental evidence from the lab (e.g., Brown et al., 2003). In this sense, Fehr et al. (2007) and I agree that repeated interactions are important if one wishes to find any effects of fairness concerns.

An important consideration, however, is that the theoretical framework proposed above highlights that such reciprocity observed in these natural field experiments can operate through M (social preferences) and W (strategic reciprocity). Even if we conclude that the mixed evidence from the naturally-occurring data combined with evidence from the natural field experiments painted a picture of an important statistical relationship, by not shutting down the operation of one channel (M or W), we are confounded in the data interpretation. Much like we would not ascribe the act of an airline providing gifts to frequent flyers as evidence of social preferences, the data from the natural field experiments can of course be driven by reputational concerns. Indeed, buttressing this argument is the fact that in my natural field experiment (List, 2006) I also sent confederates to purchase goods that could not be graded and in other cases my confederates approached sellers with whom little future interaction was expected. In both cases

a scant statistical relationship between price and quality emerged.

Controlled field experiments as well as laboratory experiments are powerful in the sense that they can *potentially* preclude that reciprocal responses (W) will lead to future material rewards, effectively isolating social preferences (M). The most common game in this spirit is gift exchange, as reported in Fehr et al. (1997). The game is a sequential prisoner dilemma game that has buyers deciding how much money to send to a seller in stage 1. In stage 2, the seller views this offer and decides whether to accept and what quality to return. The labor market setting naturally follows if wage, employer, employee, and effort are inserted in the relevant portions of the statements. The key behind this approach is that the analyst creates a one-shot environment.

In a natural field experiment testing the gift exchange hypothesis in two actual one-shot labor markets (classifying books in a library and door to door soliciting), Gneezy and List (2006) find that worker effort in the first few hours on the job is considerably higher in a “gift” treatment than in a “non-gift” treatment. This result is consonant with the bulk of laboratory evidence on gift exchange cited above.³ Nevertheless, the result in Gneezy and List (2006) wanes over time and in the long run the gift has a small and statistically insignificant effect. Such insights are consonant with results from the psychology literature is that there are important behavioral differences between short run (*hot*) and long (*cold*) run decision making.⁴ The notion that positive wage shocks do not invoke long run effects in effort levels is also consistent with

³ This result is qualitatively similar to other one-shot experiments, but these other studies typically do find statistically significant results (see, e.g., Kube et al. (2006), and Hennig-Schmidt et al.’s. (2006) field and lab treatments when employees did not know the surplus division).

⁴ In the hot phase, visceral factors and emotions might prove quite important, whereas in the cold phase immediate reactions are more carefully suppressed. In this sense, the hot/cold settings can lead to much different behaviors (see, e.g., Loewenstein and Schkade, 1999). Loewenstein (2005) reviews some of the empirical evidence on behavioral differences across cold and hot states.

Hennig-Schmidt et al.'s. (2006) field experiment (and the lab treatments when employees did not know the surplus division) and Kube et al. (2006). However, it is important to note that Kube et al. (2006) do find evidence of short and long run effects of negative gifts.

As the preceding discussion suggests, the evidence is generally mixed on the gift exchange relationship in those cases where the repeated game incentives are suppressed. Yet, one result that does appear to follow from the literature is that the impact of gift exchange in one-shot interactions on aggregate market efficiency is small. Equally as important, it is unprofitable for principals. Again, Fehr et al (2007) seem to agree, which is an important departure from the earlier literature that argued ferociously about the empirical importance of such preferences.

III. Discussion

Fehr et al. (2007) stress that the empirical literature shows that “Repeated interactions are a...powerful multiplier of the effect of fairness concerns.” While this sounds plausible, I know of no empirical evidence that unambiguously shows this result. More specifically, I am unaware of data that suggests reputational concerns by themselves do not yield the gift exchange relationships observed in the data. Moreover, as alluded to in the empirical data summary, the literature shows that in one-shot environments—both in the lab and in the field—the effect of social preferences on aggregate market efficiency is small, and trusting actions are unprofitable for principals.

Yet, I am sure that a vector of laboratory design parameters exist that yield a measurable effect of social preferences on market efficiency that is also profitable for principals. When this set of laboratory results are released, I will interpret the data cautiously because the model in Section I and the accompanying empirical evidence suggests that a wide range of factors affect the degree to which an agent's actions will exhibit pro-social tendencies, including the nature

and extent of scrutiny, small changes in the way a decision is framed, the stakes involved, self-selection of participants, and artificial restrictions on the action space or duration of play.

List (2006) presents evidence in favor of some of these conjectures. That study carries out gift exchange experiments in the lab and field which buyers make price offers to sellers, and in return sellers select the quality level of the good provided to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers. In the lab, the results mirrored the typical findings with other subject pools: strong evidence for social preferences was observed. I then carried out a second lab experiment that maintained the central elements of the gift exchange game, but in a form that was more closely aligned to the context in which sportscard trading takes place. The goods exchanged in this lab treatment were actual baseball cards whose market values are heavily influenced by minor differences in condition that are difficult for untrained consumers to detect. If social preferences are present on the part of card sellers, then buyers who offer more money should be rewarded with higher quality cards. When card sellers were brought into the lab to sell their cards, which were subsequently professionally graded, the results paralleled those obtained in the standard gift exchange game with student subjects. However, as noted above, when these same sellers were not aware that their behavior was being scrutinized, the social preferences so routinely observed in the lab were significantly attenuated in the field.

Other field generated data yield 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 monitored proves to be the critical factor influencing behavior in this study. Further, Benz and Meier (2005) combine

insights gained from a controlled laboratory experiment and naturally occurring data to compare how individuals behave in donation laboratory experiments and how the same individuals behave in the field. Consistent with the theory in Section I, they find some evidence of correlation across situations, but find that subjects who have never contributed in the past to the charities gave 75 percent of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment.

Gneezy et al. (2004) find that while behavior in a social dilemma game in the laboratory exhibits a considerable level of cooperative behavior, in a framed field experiment that closely resembles the laboratory game they find no evidence of cooperative play, even though both experimental samples are drawn from the same student population. They speculate that unfamiliarity with the task and confusion are two reasons why negative externalities are influential in the lab but not in the field. Such results are consistent with our simple model.

Overall, these results are consistent with the wealth of psychological literature that suggests there is only weak evidence of cross-situational consistency of behavior (see, e.g., Mischel, 1968; Ross and Nisbett, 1991). For instance, Hartshorne and May (1928) discovered that people who cheat in one situation are not the people who cheat in another. If this result spills-over to measurement of pro-social preferences, it means either that (a) there is not a general cross-situational trait called “other regarding,” and/or (b) the subjects view one situation as relevant to social preferences and the other as irrelevant. In either case, such insights are consonant with the model, which predicts that factors generating perceptible differences between environments can lead to important behavioral deviations.

IV. Concluding Thoughts

Akin to natural scientists, economists have employed experimental methods to lend insights into important phenomena. Recently, use of laboratory experiments to measure deep preference parameters has grown in popularity, particularly in relation to measuring social preferences. Perhaps the most fundamental question concerning this line of research is whether findings from the lab are likely to provide reliable inferences outside of the laboratory.

In this spirit, the distinct advantage of experimenting with humans becomes a potentially serious liability. The choices that individuals make depend not just on financial implications, but also on the nature and degree of others' scrutiny, the particular context in which a decision is embedded, and the manner in which participants are selected to participate. Because the lab systematically differs from most naturally-occurring environments on these dimensions, experiments may not always yield results that are readily generalizable.⁵

As I sat down after delivering my comments I was met with a nod and wink, the astute gentleman to my side kindly noted in private that experimenting with fish has some advantages after all. "Thank you Professor Solow" was the only response I could muster.

Figure 1: A Field Experiment Bridge

Controlled Data			Naturally-Occurring Data	
	Lab	AFE	FFE	NFE, NE, PSM, IV, STR
■ Lab:	Lab experiment			
■ AFE:	Artefactual field experiment			
■ FFE:	Framed field experiment			

⁵ This point, of course, applies with equal force to data generated from naturally occurring environments.

- NFE: Natural field experiment
 - NE: Natural experiment
 - PSM: Propensity score estimation
 - IV: Instrumental variables estimation
 - STR: Structural modelling
-

References

- Bandiera, Oriana, Iwan Rasul, and Imran Barankay. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, 2005 120(3): pp. 917-962.
- Benz, Matthias and Stephan Meier. "Do People Behave in Experiments as in Real Life? – Evidence from Donations." Institute for Empirical Research in Economics - Working Papers iewwp248, 2006
- Berg, Joyce, John W. Dickhaut, and Kevin A. McCabe. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, July 1995, 10(1), pp. 122-142.
- Brown, Martin, Armin Falk, and Ernst Fehr. "Relational Contracts and the Nature of Market Interactions." *Econometrica*, 2004, 72(3), pp. 747-780.
- Chen, Paul. "Reciprocity at the Workplace: Do Fair Wages Lead to Higher Effort, Productivity, and Profitability? Working paper, 2005.
- 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 Urs Fischbacher. "The Nature of Human Altruism." *Nature*, 2003, 425.
- Fehr, Ernst and Simon Gächter. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 2000, 14(3), pp. 159-181.
- _____. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 2000, 90(4), pp. 980-994.
- Fehr, Ernst; Simon Gächter, and Georg Kirchsteiger. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 1997, 65(4), pp. 833-60.
- Fehr, Ernst, George Kirchsteiger, and Arno Riedl. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics*, May 1993, 108(2), pp. 437-59.
- Fehr, Ernst and John A. List. "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 Klaus Schmidt. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 1999, 114, pp. 817-868.
- Fehr et al. (2007)

Freeman, Richard. 1987. "Does the New Generation of Labor Economists Know More than the Old Generation?" How Labor Markets Work: Reflections on Theory and Practice by John Dunlop, Clark Kerr, Richard Lester, and Lloyd Reynolds. 205-232.

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

Gächter, Simon and Christian Thöni. "Social Learning and Voluntary Cooperation among Like-Minded People," *Journal of the European Economic Association*, 2005, 3(2-3) pp. 303-314.

Gneezy, Uri, and John A. List. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, (2006), September, 74(5): 1365-1384.

Gneezy, Uri, Ernan Haruvy, and H. Yafe. "The Inefficiency of Splitting the Bill: A Lesson in Institution Design." *The Economic Journal*, April 2004, 114, pp. 265-280.

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

Karlan, Dean and John A. List. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review*, (2007), forthcoming.

Krueger, Alan B. and Alexandre Mas. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy*, 2004, 112(2), pp. 253-289.

Landry, Craig, Andreas Lange, John A. List, Michael K. Price, and Nicholas Rupp, "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 2006, 121(2): 747-782.

Levitt, Steven D., and John A. List, "What do Laboratory Experiments Measuring Social Preferences tell us about the Real World?" *Journal of Economic Perspectives*, (2007), 21 (2): 153-174.

List, John A. "Young, Selfish, and Male: Field Evidence of Social Preferences," *Economic Journal* (2004), 114(492): pp. 121-149.

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

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

List, John A. and David Lucking-Reiley. "Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *Journal of Political Economy*, 2002, 110(1): 215-233.

Loewenstein, George and David Schkade. *Wouldn't It Be Nice? Predicting Future Feelings*. In D. Kahneman, E. Diener & N. Schwartz (Eds.), *Well Being: The Foundations of Hedonic Psychology*. Russell Sage Foundation, 1999.

Loewenstein, George. "Hot-cold Empathy Gaps and Medical Decision-Making." *Health Psychology*, 2005, forthcoming.