The Economic Journal, **109** (*February*), F25–F34. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

EXPERIMENTAL ECONOMICS FROM THE VANTAGE-POINT OF BEHAVIOURAL ECONOMICS*

George Loewenstein

As a behavioural economist (an economist who brings psychological insights to bear on economic phenomena), preparing a controversy corner piece criticising experimental economics (the use of experimentation to address economic questions) is like working yourself up to enter the boxing ring against a friend. Experimental economics and behavioural economics have much in common.

Both groups can trace their origins to psychology – psychological theory in one case and experimentation in the other. Both subfields came of age in the last quarter of this century, and have gained growing acceptance within the discipline of economics as measured by almost any criterion: publications in mainstream journals, academic positions in top departments, prominence at meetings, etc.. In the current climate it is easy to forget that only 20 years ago the simple fact that an article reported experiments or discussed psychology was regarded by many editors as grounds for summary rejection.

Perhaps more importantly, many behavioural economists (BEs) use economics-style experiments, and some experimental economists (EEs) embrace psychology. Indeed, some researchers would find it difficult to classify themselves into one group or the other, and would be embraced by both groups as one of their own. There is, therefore no *inherent* conflict between the two approaches; indeed, there is good cause for synergistic coexistence. Nevertheless, there is often value in obtaining another field's perspective on what one does. Some EEs have not been particularly reticent about providing BEs with such input (see, e.g., Smith, 1991). In this essay I attempt to return the favour in a small way.

There are, in fact, many differences between the two subfields, the most important of which is one of basic orientation. BEs are methodological eclectics. They define themselves, not on the basis of the research methods that they employ, but rather their application of psychological insights to economics. In recent published research, BEs are as likely to use field research as experimentation (see, e.g., Camerer *et al.*, 1997; Babcock *et al.*, 1996). EEs on the other hand, define themselves on the basis of their endorsement and use of experimentation as a research tool. Consistent with this orientation, EEs have made a major investment in developing novel experimental methods that are suitable for addressing economic issues, and have achieving a virtual consensus among themselves on a number of important methodological

* Thanks to Iris Bohnet, Colin Camerer, Graham Loomes, Drazen Prelec, Matthew Rabin, and Richard Thaler for helpful comments. This comment was completed while the author was visiting the Center for Advanced Study in the Behavioural Sciences, which was supported by NSF grant no. SBR - 960123.

issues. As a result, EE experiments share methodological features to a much greater extent than is true of experiments conducted by psychologists. My goal in this essay is to provide some reflections, from the perspective of a psychologically-minded economist, on some of these features of EE experiments.

1. External Validity

Most empirical research methods textbooks written by psychologists begin with a discussion of the tension between internal and external validity.[1] Internal validity refers to the ability to draw confident causal conclusions from one's research. External validity refers to the ability to generalise from the research context to the settings that the research is intended to approximate. Among psychologists, experiments have the reputation of being high in internal validity but low in external validity, whereas field studies are seen as embodying the opposite characteristics. Field experiments and the quasi-experimental methods developed by psychologists are designed to increase external validity with a minimal sacrifice of the internal validity that is usually associated with experimentation.

My focus in this commentary is on external validity – the dimension on which I believe EE experiments are particularly vulnerable. Many EEs seem to believe that certain features of their experiments, such as the incorporation of market institutions, stationary replication, and carefully controlled incentives, make their experiments immune to the problems of external validity that psychologists lament in their own studies. Indeed, by naming their professional society the ''Economic *Science* Association'' EEs (deliberately, I believe) make the implicit claim that experimentation is superior on both dimensions – internal *and* external validity – to the field methods that are still more commonly employed by economists.

If the goal of EEs is to represent the behaviour of certain highly structured market settings, such as stock exchanges or auction houses, then EEs are probably justified in holding such a view. The situation is quite different, however, if one defines external validity more broadly, as I believe most EEs would like to do. The same features that make EE experiments predictive of behaviour in one class of formal markets, I would argue, actually limit their applicability to the types of economic settings that play a more prominent role in daily economic life. In the remainder of the essay, I elaborate on this point by focusing on a small number of features that are common elements of EE experiments.

2. Auctions and Markets

Many, if not most, experiments conducted by EEs incorporate market mechanisms such as double oral auctions, short-selling, etc.. Such markets have

[1] It is a curious fact that undergraduate economics majors, unlike psychology majors, rarely take courses in empirical research methods. As a result, while they may be well-tooled in regression techniques, they are typically at a complete loss when it comes to other aspects of empirical research.

remarkable efficiency properties; they even converge to equilibrium when participants are 'zero intelligence' traders (Gode and Sunder, 1993). They do so, in part, by 'disciplining' suboptimal behaviour – i.e., by reducing the wealth, and eliminating the influence, of participants who behave in a suboptimal fashion.

Efficiency, however, is not the same as high external validity. Double oral auctions or even one-sided auctions, not to mention short-selling, are rare in daily life. I do not think I have ever encountered a double oral auction outside of participating in an experiment, and the last time I participated in *any* type of auction was as a teenager when I bought a broken washing machine motor for $0.25. Most of the economic transactions that I, and probably most people, participate in, whether large or small, are notable for the lack of disciplining mechanisms. For example, if I pay too much for a car or invest my retirement savings foolishly my influence on the economy barely changes. And, unlike the stock market, in which I may be able to rely on the 'smart money' to ensure that I pay a fair price, I have no such protection when it comes to buying or selling my car or house.

The issue of external validity is of particular interest to BEs because EEs have examined the robustness, in EE markets, of many of the phenomena that BEs have identified, such as expected utility anomalies, preference reversals, non-constant time discounting, the willingness to accept/willingness to pay discrepancy (the endowment effect), fairness effects, and the winner's curse. If the patterns of behaviour observed by BEs disappeared in experimental markets – which they do not – many EEs would undoubtedly conclude that they must not be very important in the real world. However, such a conclusion would be unfounded. One could perhaps surmise that they would not be displayed by currency traders, or others who engage in repetitive transaction in a market with massive rapid feedback and short-selling. But only a small fraction of economic transactions take place in settings that have these informational or incentive features.

3. Repetition

Among EEs there is a strong, and seemingly growing, belief in the importance of repetition. Experiments often consist of a series of 'periods' in which individuals engage in the same activity repeatedly, sometimes 10 to 20 times in a row. Although such 'stationary replication' severely limits the duration and maximum complexity of tasks that can be examined, it is unquestionably a useful tool for studying the important question of how people learn in highly repetitive situations. Many EEs, however, view stationary replication not as a tool for studying learning, but as a technique for increasing external validity. The view of most advocates of stationary replication seems to be that people's behaviour at the end of a series of stationary repetitions is more representative of their behaviour in economic settings than their behaviour at the beginning.

Consider an experiment conducted by Coursey *et al.* (1987). The experiment was intended as a criticism of earlier research by Knetsch and Sinden

showing a disparity between willingness to pay for a good (a lottery) and willingness to accept compensation for giving up the same good. Some participants in Coursey *et al.*'s experiment were told they would have to taste a spoonful of a bitter substance (SOA) and an auction was conducted to elicit willingness to pay to avoid tasting the SOA. Willingness to accept compensation for tasting the substance was elicited from other subjects who were not initially made to expect that they would have to taste the SOA. The disparity between WTA and WTP began at a ratio of approximately 5, but by the end of 10 market sessions had diminished to approximately 1.6 which was not statistically significant, albeit using a low-power rank-sum test with a small number of subjects and a 0.01 significance level. The authors conclude that 'the divergence obtained in early trials of the experiment . . . may result mainly from a lack of market experience' and that 'individuals may well learn to become more rational under the pressure of a competitive market' (p. 688).

As Camerer (1996) notes, the situation that participants face in experiments of this type is somewhat akin to that of the protagonist in the film 'Groundhog Day', who repeatedly relives the same day until he 'gets it right'. Outside of this fictional film, how many people are exposed to the situation of repeatedly, and in close succession, bidding on the same good (or a bad)? Stationary replication is simply not a common feature of economic life. There are a few settings, such as working in a highway tollbooth or perhaps trading options, in which people face many highly repetitive situations in close succession, but these are probably not the most interesting or important when it come to understanding human (or market) behaviour. And they certainly are not the *uniquely* interesting case.

According to Ledyard (1995), the benefit of repetition is that it allows the experimenter 'to discover whether the data are simply the result of confusion and inexperience.' He is clearly correct that it is bad for subjects to be confused about the mechanics of the experiment. However, when the confusion reflects a lack of understanding about how to behave in a particular situation, perhaps due to the lack of relevant experience, it is by no means clear that behaviour at the end of the repetitions is more representative of actual economic behaviour than behaviour at the beginning. People may be confused about whether to obtain an advanced degree and what type of degree to obtain, about what type of job will advance their career goals, or even what those goals are. They might eventually behave optimally if they faced these decisions repeatedly and received feedback about the consequences of their choices. But they do not.

Defenders of stationary replication could argue that, although people rarely face the type of unchanging situation characterised by stationary replication, they can apply what they learn in one situation to their behaviour in other situations. For example, if one blunders in bargaining over the price of a used car it is possible that one will be less likely to blunder subsequently when bargaining over the purchase of a house. But if people are so successful at what psychologists call 'transfer of learning' across situations, it is surprising that they do not transfer such insights to experimental settings.

In fact, psychological research suggests that transfer of learning across situations is surprisingly weak. Even when subjects are explicitly informed that their experience on one task is relevant to a second task, they often learn the wrong lesson from the first task. Thus, for example, Bassok *et al.* (1995) trained subjects on a problem in which the manager of a country club randomly assigned caddies to golfers. Their task was to compute the probability that the three most experienced caddies would be assigned to the three newest members. The essential insight to solving the problem is that the answer depends on the size of the 'assigned' set (caddies) and not on the size of the 'receiving' set (golfers). Subjects were able to learn the rule that enabled them to provide relatively accurate answers to the problem. However, when the problem was changed superficially so that subjects were asked to assigning caddies to golf carts instead of golfers, 76% gave wrong answers which indicated that they had learned the wrong rule – one based on superordinate and subordinate relations (golfers are superior to caddies and caddies are superior to carts) rather than one based correctly on the assigned/received distinction.

Analogous results have been obtained by experimental economists. Consider, for example, an auction experiment conducted by Kagel and Levin (1986). Subjects who bid on an asset in a 3-person auction initially exhibited the winner's curse (by overbidding) but, after several repetitions, dropped their bids to the point where they no longer lost money. One interpretation is that subjects dropped their bids because they had learned about the winner's curse. However, when they were subsequently placed in a 6-person auction, instead of lowering their bids, as they would have done if they had learned the essential principle behind the winner's curse, they raised them. Subjects learned a rule in the first situation that reduced their losses in that situation, but did not generalise to even a subtly different situation.

Besides exaggerating the degree of learning that takes place in real-world settings, stationary replication can also affect people's preferences in a way that may or may not enhance external validity. Repetition tends to repress certain types of psychological motives, such as fairness, that may play a prominent role in early-period play. How many times can a subject get angry about someone splitting a pie unevenly? It must be acknowledged, however, that similar factors may be operative in daily life: how many trips to the wine store does it take before one forgets one's resolution to punish the French for violating the nuclear test ban treaty?

4. Context

Many experimental economists seem to view their enterprise as akin to silicon chip production. Subjects are removed from all familiar contextual cues. Like the characters 'thing one' and 'thing two' in Dr. Suess' *Cat in the Hat*, buyers and sellers become 'persons *A* and *B*', and all other information that might make the situation familiar and provide a clue about how to behave is removed. The desire to expunge context is reminiscent of a movement among

behaviourist psychologists in the middle of this century, at the peak of which some researchers conducted experiments in 'context free' temperature and sound-regulated white egg-shaped enclosures. The context-free experiment is, of course, an elusive goal. An egg-shaped cage provides the same amount of context, albeit somewhat more alien, as any other environment.

Nor would a context-free experiment necessarily be a good thing if it were possible. A major discovery of cognitive psychology is the degree to which all forms of thinking and problem solving are context-dependent, including such seemingly straightforward tasks as language-comprehension. For an example in the domain of problem-solving, consider the 'Wason four-card problem'. There are four cards, each with a letter on one side and a number on the other. The exposed faces read '*X*', '*Y*', '1', and '2'. Subjects are asked which cards would need to be turned over to test the rule: 'If there is an *X* on one side there is a 2 on the other.' Very few subjects give the right answer, which is X and 1. However, when the problem is put into a more familiar context (for example, there are 4 children from two different towns and two school districts and the rule is 'If a child lives in Concord he goes to Concord High'), a much higher fraction of subjects give the right answer. Subjects may seem like zero intelligence agents when they are placed in the unfamiliar and abstract context of an experiment, even if they function quite adequately in familiar settings. Indeed, the pervasive confusion on the basis of which Ledyard justifies the need for repetition may stem, in part, from the difficulty of explaining the experiment to subjects in the absence of any familiar contextual cues.

In addition to context effects, social psychologists have documented social contagion effects, as well as pressures to conform, than can exert a powerful influence in experimental settings. In social encounters, including laboratory experiments, most people are engaged in a constant search for cues about how they are supposed to behave. These cues can trigger off complex inferences. For example, in a study of dictator game behaviour conducted by Hoffman *et al.* (1994), dictators who made decisions behind a cardboard partition and were given detailed instructions about the elaborate measures that had been taken to ensure anonymity gave away less money than in conditions that did not ensure anonymity. The authors conclude that 'people act as if they are other regarding because they are better off with the resulting reputation. Only under conditions of social isolation are these reputation concerns of little force' (page 659). Although the effects of anonymity may result, in part, from concerns about reputation, it seems likely that what psychologists refer to as 'demand effects' also played a role. It is natural for subjects to infer from the elaborate measures taken to ensure anonymity that they are supposed to behave in a way that they would not like others to observe.

As a result of their experience with experimental games, EEs are beginning to gain an appreciation of the importance of context effects. For example, in investigations of the 'ultimatum game', fairly subtle experimental manipulations (such as whether the game is posed a matter of selling and buying or of division of a resource, and whether money offers are actually presented in cash) can have large effects on the behaviour of the players. Similarly,

defection rates in the prisoners' dilemma game are dramatically different if subjects are told they are playing the 'Wall Street Game' or the 'Community Game' (Ross and Ward, 1996). It could be argued that such results indicate that the game should be played without labels, but in such a situation subjects will inevitably apply their own labels. Unfortunately, there is no 'neutral' presentation of these games, simply a variety of alternatives, so there is no way to remove the context.[2] The goal of external validity is served by creating a context that is similar to the one in which economic agents will actually operate, not by engaging in futile attempts to eliminate context.

5. Incentives

Of all the 'rules' of EE experiments, perhaps the most stringently enforced is the use of monetery payments that are contingent on behaviour, which is seen as a means of maintaining strict control over incentives. It is difficult to think of any economics experiment published in a major journal that did not incorporate such contingent payments.

Although the use of contingent financial rewards makes good sense, EEs should not deceive themselves into believing that the use of such rewards allows them to control the incentives operating in their experiments. Even with monetary rewards, subjects are likely to be influenced powerfully by motives other than profit maximisation. Such motives include the desire to behave in an appropriate fashion, conform to the expectations of the experimenter, appear to be a smart (or at least not stupid), good, person, a winner, etc.

Some EEs seem to believe that some of these motives can be eliminated through procedures that assure anonymity. For example, Hoffman *et al.* (1994), articulating such a view, state that 'it seems unreasonable to believe that people directly consume their reputations in isolation' (p. 659). Quite to the contrary, from a psychological perspective the idea that people directly 'consume' (that is, care about) their reputations is not only eminently reasonable, but consistent with a myriad of studies (see Bodner and Prelec, 1996). For example, people behave very differently in the presence of a mirror, even when they believe that no one is observing them (Duval and Wicklund, 1972). Certainly, people do behave differently when they believe thay are being watched, but how much this is due to reputation effects is unclear. Most people

[2] Because context cannot be eliminated, experiments should never be used for the purpose of measuring individual propensities. It is tempting, but a big mistake, to think that behaviour in dictator games measures an individual's altruism, that responders' behaviour in ultimatum games measures their taste for fairness, or that behaviour in the trust game measures trustworthiness. I am aware of no evidence showing that people who give more money in dictator games contribute more to charities, are more likely to put themselves at risk to rescue a drowning stranger, or give their subway seat to the elderly or infirm. But, even if such data are collected and the correlations prove to be positive, the fact remains that it would be easy, through a suitable manipulation of context, to design a dictator study in which people would give none of their money to a stranger, or one in which most people would give all of their money away. Which of these contexts uncovers people's 'true' level of altruism? Some EEs seem to believe that they know the answer: whatever context gives results that are closest to the standard economic model.

will not pick their nose while a car is passing them on the freeway, even in a foreign country. Moreover, people may be almost as concerned about maintaining a particular *self-image* as they are about maintaining an external image.

Even if experimenters were able to eliminate motives other than profit maximisation from their experiments, it would not necessarily be a good thing insofar as external validity is concerned. Profit maximisation may be an important incentive in economic transactions, but it probably is not the most important. Anyone working in a business setting, for example, can attest to the power of social comparison. Most academics seem to be motivated more by ego than by the size of their salary, and academics probably are not exceptional in this regard. Even when people do care about money, their degree of concern is often remarkably unrelated to the amount of money involved. Small amounts of money can gain momentous significance under certain circumstances – for example, if one gets an undeserved speeding or parking ticket, or if one is unfairly denied a small year-end bonus. Much larger changes in wealth, such as those which result from a change in stock prices, can leave one remarkably unmoved.

Moreover, monetary incentives interact with nonpecuniary incentives in ways that are poorly understood. For example, in a tournament situation, there is likely to be an extreme discontinuity between no difference between the winning and losing prizes and even trivially small differences. Any difference in prizes that tips players off as to who won and who lost is likely to unleash a tremendous amount of motivation that will be relatively insensitive to further increases in the difference. Although the use of monetary incentives in an experiment probably rarely decreases external validity, given that they are probably not the most important source of motivation in daily economic life, their contribution to external validity is likely to be minimal.

6. Internal Validity

Although my focus has been on the issue of external validity, it is worth noting that EE experiments, like all empirical research, also face threats to internal validity. Many of these threats to internal validity result from a common failure by EEs to assign subjects randomly to different treatments. For psychologists and BEs, random assignment to treatment groups is *the* single most critical measure for achieving high internal validity. As a BE, I have been repeatedly shocked by EEs who, unapologetically, make direct comparisons between treatment groups run at different times (even sometimes in different years) and with different populations of subjects. The problem is exacerbated by the fact that observations are often not independent of one-another because subjects provide multiple observations and interact with one-another.

As virtually all introductory empirical research textbooks discuss, random assignment eliminates, in one fell swoop, a long list of threats to internal validity such as, selection, maturation, history and regression to the mean. If an experimenter conducts a study of behaviour in dictator games and compares one treatment collected in one class on Monday with another collected

in a different class on Friday, any observed difference may be due to the experimental treatment, but it is equally plausible that it resulted from differences in the composition of the two classes, important news events that took place during the week, or course materials that subjects were exposed to during the week. For the BE, running an experiment without random assignment defeats the whole purpose of the enterprise.

7. Concluding Comments

In more than thirty years of productive research, EEs have developed some extremely compelling experimental conventions. For example, EE methods of ensuring incentive compatibility, the discouragement of deception, and the practice of reporting methods in sufficient detail to allow for replication, are all worthy of broad emulation. However, despite (or even partly as a result of) their incorporation of markets, repetition, and incentives, EEs have not, in my opinion, been able to avoid the problem of low external validity that is the Achilles heel of all laboratory experimentation. EE experiments have high external validity if they are intended to represent the behaviour, and consequences of that behaviour, of people operating in highly structured markets. They are much less well suited for testing predictions about the economic consequences of individual behaviour in the 'real world,' including the real world of decentralised markets.

Whatever the external validity of their experiments, EE-style experiments are ideal for examining individual behaviour under conditions of varying incentives, opportunities for learning, interpersonal interactions, etc.. Given that BEs have proposed some of the most novel and provocative hypotheses about individual behaviour, BE may well be the single best application of EE methods.

Carnegie Mellon University, Pittsburgh

References

Babcock, L., Wang, X., and Loewenstein, G. (1996). 'Choosing the wrong pond: social comparisons that reflect a self-serving bias.' *Quarterly Journal of Economics*, vol. 111, pp. 1–19.

Bassok, M., Wu, L. L. and Olseth, K. L. (1995). 'Judging a book by its cover: interpretative effects of content on problem-solving transfer.' *Memory and Cognition*, vol. 23, pp. 354–67.

Bodner, R. and Prelec, D. (1996). 'The diagnostic value of actions in a self-signaling model,' MIT working paper.

Camerer, C. F. (1996). 'Rules for experimenting in psychology and economics, and why they differ.' In (W. Albers, W. Guth, and E. Van Damme) *Experimental Studies of Strategic Interaction: Essays in Honor of Reinhard Selten*. Berlin: Springer-Verlag.

Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. (1997). 'Labor supply of New York City taxi drivers: one day at a time'. *Quarterly Journal of Economics*, vol. 112, pp. 407–41.

Coursey, D. L., Hovis, J. L. and Schulze, W. D. (1987). 'The disparity between willingness to accept and willingness to pay measures of value.' *Quarterly Journal of Economics*, vol. 102, pp. 679–90.

Duval, S. and Wicklund, R. A. (1972). *A Theory of Objective Self-awareness*. New York: Academic Press.

Gode, D. K. and Sunder, S. (1993). 'Allocative efficiency of markets with zero-intelligence traders: market as a partial substance for individual rationality.' *Journal of Political Economy*, vol. **101**, pp. 119–37.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). 'Preferences, property rights, and anonymity in bargaining games,' *Games and Economic Behaviour*, vol. 7, pp. 346–80.

Kagel, J. and Levin, D. (1986). 'The winner's curse and public information in common value auctions,' *American Economic Review*, vol. 76, pp. 894–920.
Ledyard, J. O. (1995). 'Public goods: a survey of experimental research.' In (J. H. Kagel and A. E. Roth, eds.) *Handbook of Experimental Economics*. Princeton: Princeton University Press, pp. 111–94.
Ross, L. and Ward, A. (1996). 'Naive realism: implications for social-conflict and misunderstanding.' In (E. S. Reed, E. Turiel and T. Brown). *Values and Knowledge*, Mahwah, NJ: Lawrence Erlbaum Associates.
Smith, V. (1991). 'Rational choice – the contrast between economics and psychology.' *Journal of Political Economy*, vol 99, pp. 877–97.