Experiments in Economics: Testing Theories vs. the Robustness of Phenomena*

Francesco Guala† and Luigi Mittone‡

1. Introduction

There's a commonly held view about the role and goals of experimental economics that we think is greatly mistaken. According to such a view, the goal of laboratory experimentation is to test economic theories (or models). This view is not restricted to experimental economics, and therefore economists should not be particularly blamed for it. It is part of a popular view of science, and as such informs the rhetoric of many disciplines. Moreover, it has gained philosophical legitimacy by being systematised in the so-called 'Standard View' that dominated philosophy of science for the best part of the last century. Yet, as we shall argue, it is mistaken and exerts a bad influence on experimental economics.

2. The Theory-Testing View

It is widely believed that science is about finding the 'best' theories for the explanation, and possibly prediction and control, of real-world phenomena. It is also widely believed that experimental data play a crucial role in the selection of theories, by eliminating weak candidates and corroborating promising ones. This view, if stated in such vague terms, sounds roughly correct. But as soon as it is articulated and made more precise, it turns out to be problematic. According to the *theory-testing view* of science, as we shall call it, theorists propose, and experimenters control (and, sometimes, dispose). This view has been systematised in the so-called 'Standard-View' in the philosophy of science during the first half of the twentieth century.[1] According to the Standard View theories are sets of sentences including at least one candidate for a law of nature, i.e. a universal statement of unrestricted domain of applicability, expressed in conditional form ('For all entities x, if x has property P then it also has property Q'). Candidate laws, along with 'bridge laws' connecting them to phenomena and statements of initial conditions, are used to derive predictions about particular events. After a prediction has been put forward, observed data can be used to test the theory. The data ideally should provide one of two answers: if they are consistent with the prediction (i.e. if the predicted phenomenon has been observed), our confidence in the truth of the theory is increased. If in contrast the data are inconsistent with the prediction, our confidence is diminished.

---

* Parts of this paper derive from a seminar Francesco gave at the University of East Anglia in October 2001. The experiments were run at CEEL with the crucial help of Marco Tecilla. We thank John Broome, Riccardo Boero, audiences at East Anglia and at the 2002 Experimental Economics Workshop in Siena for their insightful comments and critiques. The usual caveats apply.
† Cognitive Science Laboratory, University of Trento, and Centre for Philosophy of the Social Sciences, University of Exeter; f.guala@ex.ac.uk.
‡ Department of Economics and CEEL, University of Trento; mittone@gelso.unitn.it.
[1] Cf. e.g. Popper (1934), Nagel (1961), and Hempel (1965).

The standard view, to be sure, is entirely general and makes no principled distinction between experimental and non-experimental tests. The practical difference is that in the laboratory we can set up and control more tightly the initial conditions, and thus make sure that the right conclusions can be drawn from the observed data. Imagine a fictional astronomer trying to use Newton's theory to predict the orbit of a planet. He observes the position of the planets in the Solar system at time t_1, and on the basis of the laws of classical mechanics and a set of boundary assumptions derives a prediction about the position of the planet at t_2. But of course he may misidentify the position of the planet, or of the Sun or some other planets (his telescope may be faulty). Or some comet, black hole, or another unexpected entity may enter the scene, mess up the initial conditions, and hence the prediction. When he finds out that the planet is not where it was supposed to be at t_2, the scientist will not be sure whether to blame the theory or one of these unexpected interferences or changes in the initial conditions of the system under observation.

This problem – known as the 'Duhem-Quine problem', from the physicist and the philosopher who introduced it in the philosophical literature – is particularly critical in the social sciences, where possible disturbing factors are numerous and 'background' circumstances tend to change frequently. The laboratory may reduce the practical import of the problem quite drastically, by letting the scientist set up the initial conditions, shield the system from external influences, and monitor it carefully so as to avoid unexpected disturbances.[2] Most experimental economists, we reckon, embarked in this new programme precisely with the hope of finding a more effective methodology than the traditional non-experimental approaches, and therefore improving the empirical basis of economic science. Yet, these hopes must face some challenging problems, which we discuss in the next section.

3. Problems with the theory-testing view: external validity and the vagueness of economic models' assumptions

Alongside with impressive confirmations of the predictions of some economic models, experimental economics has also produced stunning refutations of some basic principles of economic theorising. This is to be expected from a genuinely empirical discipline, but reactions to these findings have been varied and conflicting. Some economists have welcomed the experimental anomalies as the ultimate proof that the fundamental principles of mainstream economics are flawed and need drastic revision. Others remain unconvinced and rebut by means of a simple but powerful argument: economic models are supposed to be applicable to *real* economies, not to the 'artificial' conditions implemented in the economic lab. This critique raises a problem that is well-known among experimental psychologists and other social scientists, but is surprisingly little discussed in economics: the problem of *external validity* of experimental results.[3] The critics, in other words, claim that we should not believe that experimental results can be transferred to 'real-world' economies.

[2] The Duhem-Quine problem is logically-speaking unescapable, because we can never be sure that the right controls have been implemented. But that's life; what really matters is that there are practical ways in science to reduce the problem's impact.
[3] When it *has* been discussed, this problem has been referred to by experimental economists as the problem of 'parallelism'. Cf. e.g. Smith (1982), Wilde (1981).

Experimental economists are annoyed by such remarks, especially when they are unexplained (i.e. they don't specify what the difference between laboratory and real-world economies exactly amounts to), and are put forward by people who have no deep understanding of laboratory practice. Yet, there is something in the above critique, which points to a basic flaw in the theory-testing view and should be taken seriously. The theory-testing view assumes that science is a game with two players (see Figure 1). On the one hand, we have theory (which in economics takes the form of sets of models), on the other reality. Theory is used to ask questions, the real world answers these questions, and the theory is modified in order to take the answers into account.

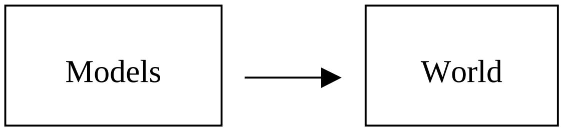

Figure 1

But this view is too rough. In particular, it presupposes that *everything* that takes place on the right-hand side is relevant for the appraisal of economic models. This is just what the critics of experimental economics deny: economic theory is aimed at explaining (and help in predicting and controlling) only a certain *part* of reality, and what happens in economic labs does not necessarily (or primarily) fall into the intended domain of application. There are two standard replies to this challenge, which unfortunately are both unconvincing.

(a) Economic theories are general in scope of application. This view has been defended, among others, by Charles Plott (1991, p.905): "[economic] models are general models involving basic principles intended to have applicability independent of time and location". Therefore, the argument goes, behaviour in laboratory settings falls automatically within the domain of economic models, and whatever is observed in there is relevant for the appraisal of economic theory. This view has a very prestigious philosophical pedigree. It belongs, in fact, to the Standard View of theories introduced in section 2 above. The Standard View, however, was put forward with physics in mind, and in fact the requirement that theories should be general in scope of application makes more sense in physics than in other scientific disciplines.[4] As a matter of fact, most economic models describe mechanisms and phenomena embedded in fairly specific institutional settings. It is pretty obvious, for example, that the laws of supply and demand, or the mechanisms of market clearing, work only when the 'right' conditions are in place[5], but do not take place in, say, a centrally planned economy. The argument does not fare better if we shift the focus to more fundamental principles like the rationality assumptions of expected utility theory. It is now commonly held that the sort of rational behaviour postulated in economic models takes place only against the 'right' background conditions – in transparent settings

[4] Although it is not uncontroversial even in the realm of physics: cf. Cartwright (1999).
[5] Part of the contribution of experimental economics, in fact, has been to identify some of the conditions that make such phenomena and mechanisms possible.

with learning and repetition, for instance (Plott 1995; Binmore 1999), but sometimes fail in less than ideal environments.

One might reformulate the argument in normative form: perhaps economic models are not general in scope and application, but they *should* be. The generality requirement, in other words, would be a desideratum, an ideal that guides the development of science.[6] The problem here is that there are good reasons to believe that the ideal cannot be fulfilled in disciplines like economics. Unlike physics, which is concerned with the discovery of the most basic properties of matter, sciences like economics (or, for that matter, biology, psychology, and so on) investigate reality at a non-fundamental level. It is highly likely that entirely general laws simply do not exist at such level, because the entities and properties of economic science (preferences, expectations, consumers, firms, markets) and the relations holding between them are non-fundamental in character. Most 'laws' in the social sciences are *ceteris paribus* in character, and the ceteris paribus clause covers conditions and factors that go well beyond the boundaries of economics. Of course one may try to overcome such boundaries, to include all the factors and conditions that are sufficient for the instantiation of economic effects; but then economics would become something very different from what it is now, probably closer to psychology and neurophysiology. At any rate, this sort of reduction is a long way off our present capacities, we need a science of economic phenomena in the meantime, and to set unrealistic goals may do more bad than good by taking economists' attention away from what can be really achieved.[7]

(b) Economic experiments should mirror the assumptions of models. According to the second argument experiments are devised to teach us about theory, and one should not worry about the real world at all. If the theory is simple, the experiment must be simple. If the theory is too simple to be applicable to reality, that is a theorist's problem rather than an experimental one (Smith, 1982, p. 268). This argument has the rhetorical advantage of shifting the burden entirely on theoretical economists.[8] But, again, it does not work. An important implication of the argument outlined in the last section is that it is unreasonable to assume that economic models can include a full description of the conditions for their application. In other words, economic models usually do not (and cannot) carry their domain of application written in their assumptions.[9] Models provide at best a partial indication of the sort of circumstances in which they are supposed to work, and part of the skill of the applied scientist consists precisely in the identification (and, sometimes, instantiation) of the implicit conditions for their application.

This, by the way, is true across *all* science: experimental physicists are no less obliged than economists to use their imagination and skill in order to create the right conditions for the instantiation of a given model or the replication of a phenomenon.[10]

---

[6] This position is strongly defended by the philosopher Karl Popper (1957).
[7] Notice, by the way, that the reduction of economics to neurophysiology or physics is not a very attractive goal for a social scientist. For a more detailed defence of the argument in the main text see Fodor (1974).
[8] Which, curiously, are sometimes very sceptical about experimental economics.
[9] This is one rationale behind Milton Friedman's (1953) famous thesis on the inevitable 'unrealisticness' of the assumptions of economic models.
[10] Cf. for example Collins (1985) and Gooding (1990).

The view according to which theory-testing is just a matter of following the instructions of a theoretical model, a machine-like procedure guided by theory from the beginning to the end, is just a myth. These considerations cast serious doubts on the view that to shift the focus from real economies to theoretical models was the greatest innovation of experimental economics. Plott (1991, p. 906) goes as far as to say that an experiment "should be judged by the lessons it teaches about the theory and not by its similarity with what nature might have happened to have created". In contrast, we think that to view scientific methodology as a play with just two characters (abstract theoretical models and experiments) is highly misleading. In the next section we sketch a view that takes the problem of external validity seriously, and casts the role played by experiments and models in a different light.

4. Experiments as mediators

When we fund medical research, say, on the effects of a new drug, we expect eventually to receive back some result that is relevant for us, human beings. We would be disappointed if in the end we were provided with a detailed study on the effects of a drug on mice, guinea pigs, monkeys, etc., but nothing at all on its efficacy on the human form of that disease. The same applies to economics. It would be embarrassing, we think, to admit that what experimental economists learn cannot be extended outside the laboratory walls. Thus, eventually, scientific results need to travel all the way to the 'real world'. The picture of science we endorse is represented in Figure 2 below.

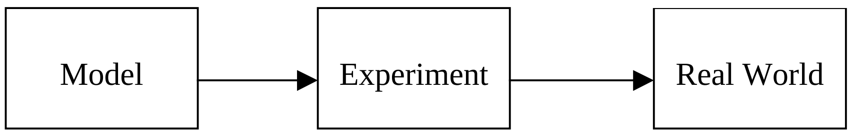

Figure 2

According to such a view, experiments are just an intermediate step in the route from pure theory to real-world economic phenomena. They are 'mediators', in the sense that they help bridging the gap between models and their intended domain of application.[11] The worse aspect of the theory-testing view is that it induces to think of models (theory) and experiments (data) as two very different things, where in reality they are not. Both models and experiments should be thought as *systems* that we are studying. Models can be abstract or, less frequently, concrete. Experimental systems are obviously more concrete than models, and closer to the intended domain of application because they include features that are held in common with the systems we are eventually interested in understanding (the real-world economies). But they are not the target system, and to move from experiment to target requires an inference. Just like the model-to-experiment inference, the experiment-to-target inference is inductive or ampliative (what we know about X does not allow us to derive deductively the properties of Y), and hence fallible.

[11] See also Guala (1998, 1999). The term 'mediators' has been borrowed from Morgan and Morrison (eds. 1999). The thesis defended here is close to the so-called Semantic View of theories; cf. e.g. Giere (1988). For a similar view of modelling in economics, cf. Sugden (2000).

There is quite a lot of tacit and explicit knowledge about how to bridge the gap between models and experiments -- or how to solve the problem of '*internal validity*', to follow the terminology of experimental psychology. At a fairly abstract level, internal validity is achieved by repeatedly testing single causal hypotheses. If you believe that effect Y may be due to factor X, you run an experiment where X is varied and other possible factors are kept constant or eliminated altogether. The same can be done for other factors, until possible interferences, background factors etc. have been checked one by one.[12] To put it sharply: *experimental economics is for hypothesis testing, not theory testing.* The prima-facie appeal of the theory-testing view derives from overlooking this basic distinction. An experimental hypothesis is usually concerned with local circumstances, singular factors, very specific sources errors, and may (but need not) be suggested by theory. Experimenters formulate hypotheses all the time, about the incentives, the experimental design, the properties of the error term. Theories explain by unifying, whereas experimental hypotheses do not. Theories are supposed to be of general applicability, whereas experimental hypotheses concern specific features of specific laboratory systems. Often they are inspired by empirical data, or by intuition, and in any case the rejection of specific hypotheses has no direct consequences regarding the acceptance of theoretical models. That is not the main job experiments, as we have seen.

At a more concrete level, the task of establishing internal validity of an experimental result depends on a lot of context-specific knowledge and techniques. Take any textbook on experimental methods in economics (e.g. Friedman and Sunder 1994, or Davis and Holt, 1993) or in the social sciences in general (e.g. Frankfurt-Nachmias and Nachmias 1996) and you will find several pages on how to control preferences, how to rule out undesired effects, and eventually how to infer from observed statistical frequencies to causal relations between properties of the experimental system. Of course much is left to intuition and the creativity of the experimenter, but the basic strategies are well known. In contrast, we find very little explicit advice about how to bridge the second gap, the one from experiment to reality (external validity problem).

5. Mimicking the target

There has been a lot of excitement recently about the applicability of game theory to solve complicated problems of market regulation. The most famous application, the auctions for third-generation mobile phone licences, were as a matter of fact a product as much of game theory as of experimental economics. Since the story has already been told elsewhere,[13] we shall not repeat it here. For our present concerns we just need to notice that the construction of the Federal Communication Commission auctions is an instance of direct attack on the problem of external validity: in this case the 'real world' itself (the market for portable communication systems) has been shaped so as to fit the experimental prototypes in the lab. Again, there is nothing peculiar here – this is just what happens with most technology, from space probes to the TV sets in our homes.

[12] See Mayo (1996) for a sophisticated analysis of experimental inference.
[13] Cf. Plott (1997), Guala (2001).

But economists usually are not so lucky: they can rarely shape the world as they wish. It is more common therefore to follow the opposite strategy. Instead of shaping the world on the experiment, the experiment is changed so as to resemble the real world in as many ‘relevant’ respects as possible. In order to do this, of course, we need to know what sort of system we want to extend our results to. We need, in other words, a specific *target*. There are various examples of this strategy: to remain in the field of auction theory, experimental research on the winner’s curse started precisely with the aim of replicating a target phenomenon, allegedly observed in the auctions of the Outer Continental Shelf.[14] The advantage of this strategy is that it makes the external validity problem tractable. If you observe phenomenon X in the laboratory, but you are not sure about its robustness to real-world circumstances *in general*, it is difficult to tackle the problem constructively. You can do much more if you know exactly the sort of circumstances you want to export your results to: in this case you can look for *specific reasons* why the result may not be exportable. These reasons will usually take the form of some dissimilarity between the experimental and the target system. Thus, the obvious way to proceed is to modify the experiment to include the feature of the target that could be responsible for the external validity failure, and see whether it does in fact make a difference or not. For example: if you think that real businessmen are different from students, use businessmen in your experiments; if you think that an ascending auction is different from a descending auction, use the former in your experiment; and so on.

Notice that to add realistic details to an experiment implies increasing difficulties in the interpretation of the experimental results. In this sense there is a clear trade-off between internal and external validity: the simpler the experimental environment, the easier it is to identify the cause(s) responsible for a given phenomenon or effect. This is why experimenters like sober designs, where the subjects are engaged with very abstract tasks. This is also why, we think, experimenters like to replicate quite literally the idealised assumptions of theoretical models: because it is one way of achieving simplicity of design.

Unfortunately the real world usually isn’t sober. By ‘taking away’[15] complications we also move away from the target. Choosing ‘to go abstract’ rather than ‘concrete’, therefore, has unpleasant implications in any case, if we remain focused on a *single* experiment. But the negative consequences of this trade-off can be partly neutralised by performing a series of experiments. Notice that internal validity is logically and epistemically prior: it doesn’t make much sense to ask whether an experimental result can be exported to a given target system, if you are not even sure what you result is. Thus, it is quite common to try first to achieve an understanding of a phenomenon (or entity, institution, etc.) in a fairly abstract, ‘bare-bone’ version. Then, the experiment is progressively ‘concretised’ in order to become more and more similar to the target of interest. This is, by the way, what happens in biomedical science: in order to test a new drug experimenters start with animal models,[16] move on to human beings in

[14] Cf. Kagel and Levin (1986) and Guala (1998).
[15] This is the etymological sense of the term ‘abstraction’: to take away, from the Latin *ab-trahere*.
[16] We are simplifying drastically here: to find out which animals are ‘right’ for which kind of investigation is not a trivial matter. Notice that animals are used for ethical reasons, not only for the dangers involved, but also for the possibility of controlling, manipulating and simplifying the environment they live in (temperature, diet, social relations, etc.). On animal models in biomedical science, see LaFollette and Shanks (1995) and Ankeny (2001).

‘ideal’ experimental settings, and conclude with so-called efficacy trials with patients in more realistic conditions.

6. The discovery of new phenomena

So far, so good. But the auction case is fairly special. Although there are features of real-life auctions that cannot be mimicked in the laboratory (think for instance at the huge sums of money spent in the markets for mobile phone licences), the experimenter can surely go a long way towards replicating features of the target system. In most other cases of experimentation this can be done to a much lesser extent. Still, we feel that even in such cases we have the possibility of learning something of wider applicability than a mere laboratory game. But what *exactly* can we learn? Let us examine a concrete example.

One of us faced this question while working on a series of experiments on tax evasion.[17] Tax evasion is a particularly tricky area of investigation, because it is commonly believed that experimental subjects tend to engage in behaviour that has little to do with the target phenomenon.[18] There are problems of scale, once again (the sums involved are small compared to real tax payments), but also of game-like behaviour (subjects tend to play with the experiment, rather than take it seriously), of absence of social incentives (your family and friends don’t know, and don’t care, if you are busted by the experimental taxman), of general unrealisticness (no lawyers, no accountants), and many others. However, from a theory-testing perspective it is not clear that these should all count as flaws of the experiment. Standard microeconomic theory models tax evasion basically as a lottery, where the agent has a given probability of being audited and therefore fined, and the utility varies only over money (the utility of keeping income instead of paying the tax, and the disutility of paying the fine). Social blame, shame, etc. do not enter the picture, nor do other social norms and institutions. But as we have already argued, the theory-testing view is wrong, and this case just provides more evidence that it is.

In the experiment ran at CEEL subjects were provided with a set of parameters (income, the amount of tax to be paid, the probability of being audited) and were asked to decide how much tax to pay effectively in each round. The experiment was concerned with dynamic choice, and thus was extended over a series of 60 rounds. The experiment was originally intended to test the effect of tax yield redistribution: when the yield is redistributed, the tax experiment becomes a sort of public goods experiment, with evasion basically analogous to free riding. Given that (contrary to standard theory) cooperative behaviour has been extensively observed in public goods experiments, by analogy the redistribution of yield may have the effect of reducing the rate of evasion behaviour. In fact this conjecture was confirmed by the data.[19]

This result, however, is just the beginning of the story: as we said, theory-testing is not the main nor the most easily attainable goal of experimental economics. A most

---

[17] The experiments discussed in this part of the paper are described in more detail in Mittone (2002).
[18] Cf. for instance Webley et al. (1991, pp. 39-47).
[19] Also other aspects of subjects’ behaviour cannot be explained by means of standard economic theory, unless some very peculiar assumptions are made concerning risk-attitude. For example, the highly erratic path of tax payments is compatible either with quickly changing attitudes towards risk, or with risk neutrality and random behaviour. Cf. Mittone (2002) for more details.

valuable feature of laboratory experimentation, one that makes it almost unique in the field of social science, is that it sometimes leads to *the discovery of new, unexpected phenomena*. Moreover, unlike field observation, laboratory work usually allows the demonstration that (1) the phenomenon in question is real and not just a spurious regularity or an artefact of statistical analysis; (2) that it is robust to changes in background factors.

A number of phenomena discovered in the lab have passed tests (1) and (2), and are now widely discussed in the economic literature: take violations of rationality such as the Allais paradox or preference reversals, but also the efficiency properties of double oral auctions, the decay of contribution in public goods experiments, and so on. Here we shall discuss just two phenomena that emerged from the tax experiments: we shall call them the ‘bomb crater’ and the ‘echo’ effect.

7. Bomb craters and echo

They say that troops under heavy enemy fire hide in the craters of recent explosions, because they believe it highly unlikely for two bombs to fall exactly in the same spot at short time-distance. Something similar seems to happen in the tax experiments: immediately after each audit, tax payments fall sharply (i.e. evasion increases). The ‘bomb crater’ effect is represented in Figure 3.

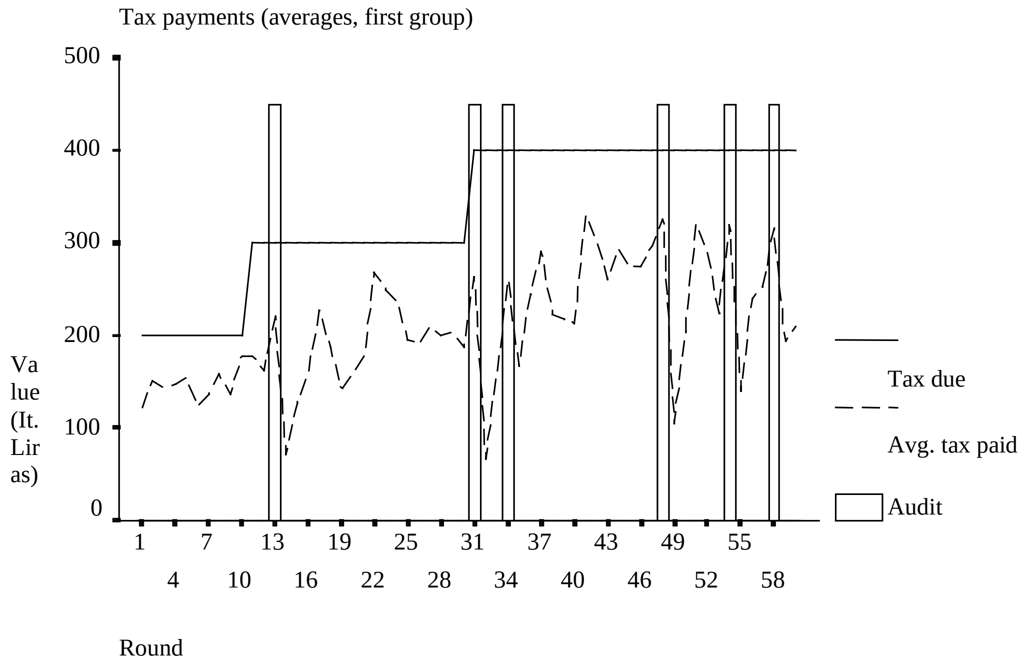

Figure 3

When a phenomenon is observed for the first time, one wants to know how robust it is to changes in experimental conditions. The bomb crater effect turns out to be remarkably robust: it persists under changes in the methods used to inform subjects of the probability of being investigated, under changes in the fiscal audit system, and is not influenced by the tax yield redistribution. Moreover, it seems to crop up in a number of other situations that have nothing to do with taxes and audits. Figure 4 represents a game invented by Luigi Mittone and tried repeatedly in the experimental lab at the University of Trento. Two players move sequentially one of three concentric wheels. The wheels can be moved only counter-clockwise, 90 degrees at a time. Player A's payoffs are given by the sum of the figures which end up in the north-western quadrant, whereas Player B receives the payoffs in the north-eastern quadrant. The game is essentially a coordination game, which can be used to investigate the computational capacities of experimental subjects (how many rounds can they anticipate, in reasoning about the game?) and their ability to communicate with the other player the existence and willingness to follow a given strategy.[20]

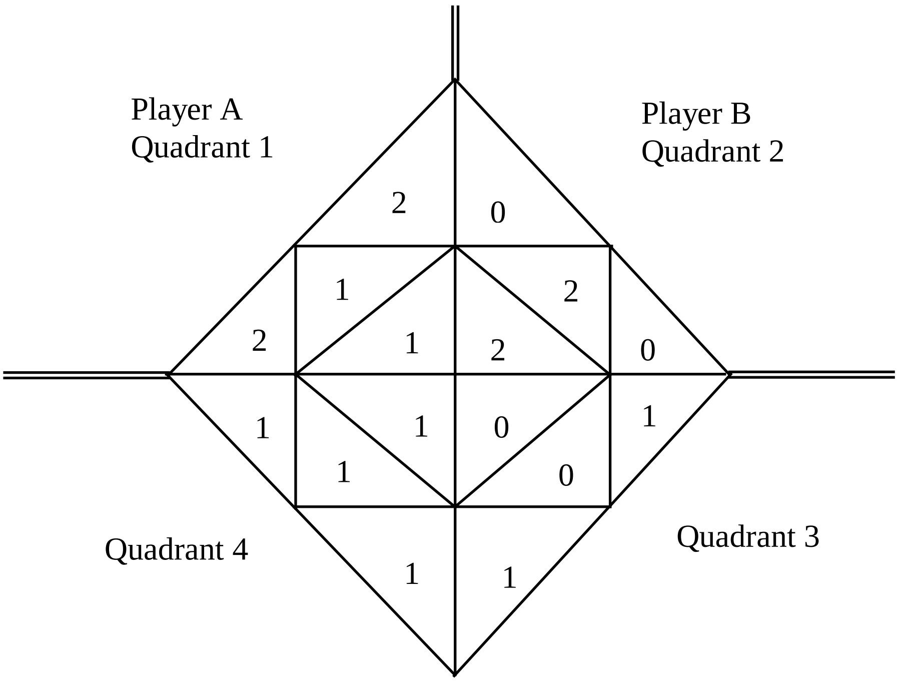

Figure 4: The 'wheel of numbers'

Now, it is possible to use this completely abstract, decontextualised game to observe the bomb crater effect in action. In order to do that, it is sufficient to add a random device, selecting every now and then one of the three wheels. If the selected wheel has just been moved by a player, that player gets a payoff of zero. Laboratory data show that the mere presence of a random device of this sort induces players to make a number of irrational moves that are normally avoided in this game. Many players tend to move the wheel that has just been selected by the random mechanism, consistently with the 'bomb-crater fallacy', even though that move is clearly dominated by an alternative one. Table 1 reports the results of this experiment.

[20] It is possible to prove the existence of several Nash equilibria in this game, but given the focus of this paper and the complex structure of the game we shall bracket such theoretical matters here.

Group	No. of rounds: chosen wheel = selected wheel	No. of rounds: chosen wheel = selected wheel, even though dominated	Percentage of rounds: chosen wheel = selected wheel, even though dominated
1	19 out of 49	6 out of 19	35.5
2	17 out of 49	7 out of 17	41.1
3	18 out of 49	8 out of 18	44.4
4	17 out of 49	6 out of 17	35.2
5	18 out of 49	9 out of 18	50.0

Table 1: The wheel of numbers

The relevant data are in the two columns on the right-hand side: between 35 and 50% of the times, the random device 'attracted' a player towards a dominated move. What does this tell us about the generality of the bomb crater effect? The most plausible answer is that we are dealing with a fairly common bias, which tends to arise whenever subjects have to do with probabilistic reasoning of this kind.[21] To establish robustness is to establish a sort of generality, to a set of situations that are similar to the ones in which the phenomenon has been observed. Robustness invites 'generic' confidence, in the sense that it is no evidence that the phenomenon will occur in *all* circumstances, and provide no precise indication of the situations in which it will occur and those in which it will not. Stylised facts from the real world invite caution: there are reasons to believe, for instance, that erratic behaviour such as the one observed in the tax experiments (a variance exacerbated by the crater effect) may not arise in real-world circumstances. Some governments take erratic tax payments as indicators of possible evasion, and therefore check erratic tax payers more often than the others. This strategy (if known to tax payers, which is an empirical hypothesis) may be enough to discourage the bomb crater effect.

The 'echo' effect is perhaps more promising from an external validity viewpoint. After an audit, for some subjects and under certain circumstances, evasion remains high for a few rounds – as if the falling of the bomb produced reverberations. This prompts the consideration that the opposite may also be true: if we managed to induce more law-abiding behaviour, the effect might be relatively enduring. If audits had long-lasting psychological effects, then, such resilience could be exploited for policy purposes. In order to check this hypothesis Mittone (2002) ran two separate sessions, in which audits were confined to either the early (Figure 4) or the late rounds (Figure 5) respectively. Repeated auditing seems to have quite a strong effect in inhibiting evasion, when it takes place in the first half of the experiment. In contrast, if subjects experience a long period of unpunished evasion at the beginning of the experiment, even a series of audits do not manage to raise the average level of tax payment. Subjects seem to become more risk-takers, and apparently it takes time for them to revise their attitude. Again, this sort of attitude is robust to changes in design conditions. Moreover, we can think of several examples from real life which seem to support the generalisability of the 'punishment' effect. In Italy, for example, where police officers are well known for their inconsistent attitude towards fining, car drivers seem to proceed at lower speed on roads upon which the police have

[21] This bias is related (perhaps a special case) of the so-called 'gambler's fallacy', i.e. the tendency to over or underestimate probabilities based on a limited sample of events. This sort of fallacy has been elicited by psychologists in experiments on animals and human subjects since the 1930's (cf. Brunsvik 1939; Brunsvik and Herma 1951; Jarvik 1951).

consistently focused across a short period of time. Something like this might happen with tax audits: repeatedly auditing an individual or group of people may cause a robust reduction of evasion for quite a long time after the event.

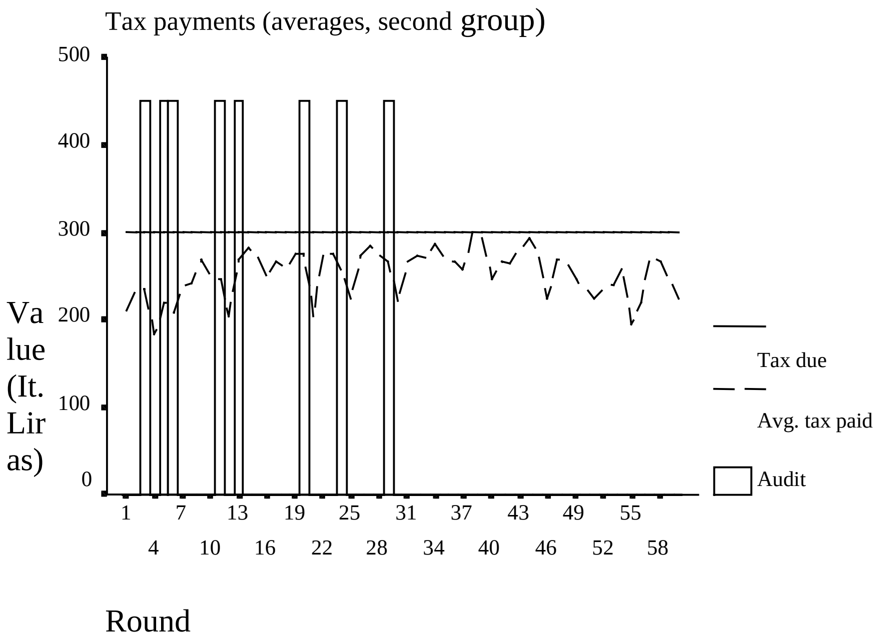

Figure 5

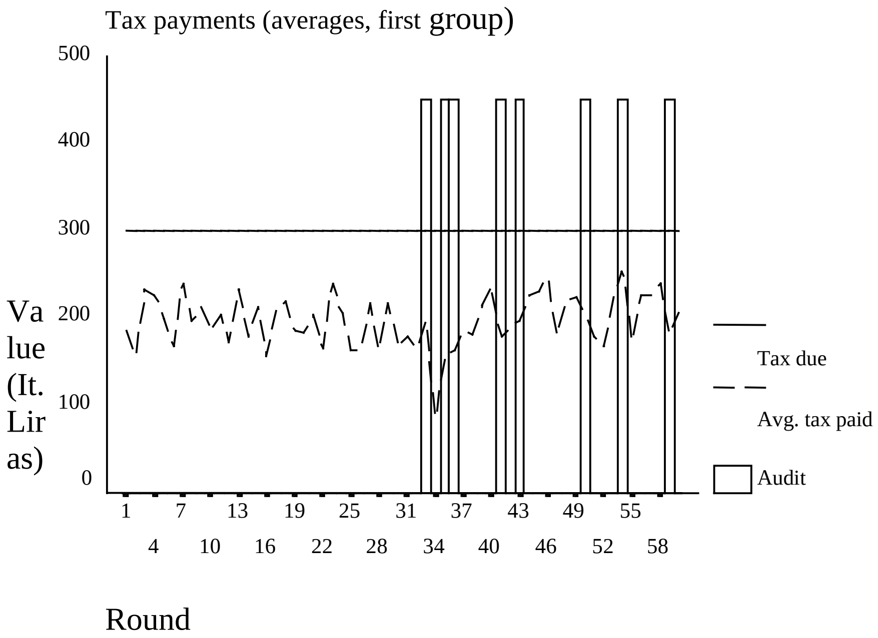

Figure 6

8. Experimental residuals

The echo effect is probably a *strong* phenomenon, which promises to be applicable to a wide range of non-experimental circumstances. Why exactly are we inclined to say so? Part of the answer lies in its unexpected character. The bomb crater and the echo phenomena were noticed 'post hoc', while analysing data collected to test a different phenomenon (the effect of tax yield redistribution). This fact, quite paradoxically, improves rather than affects for the worse its credentials as a non-purely experimental phenomenon. The underlying reasoning goes as follows: an experiment is usually designed to test the effect of a series of factors or independent variables (X_1, X_2, …, X_n)[22] on a dependent variable (Y). Usually, the experimenter tries to design an experiment such that no other factor besides $X_1,\ldots,X_n$ is likely to have an influence on Y. (This is why, as we have already noticed, abstract designs facilitate experimentation.) Then, one factor (say, X_1) is varied while the others are kept constant, and the procedure is iterated for the other $X_2,\ldots,X_n$.

The list of potentially relevant X_i may come from theory, from previous experimental results, or just from common sense. In the original tax experiment described above, the main variable at stake was tax yield redistribution, which in theory should make no difference but in practice (given the evidence from public goods experiments) is likely to have some. The idea of having redistribution can also be seen as an attempt to 'import' into the experiment some real-world features and thus increase the

[22] Of course in some cases these may be a singleton, when we focus on just one factor.

external validity of the experiment (along the lines sketched in section 5 above). But this attempt is convincing only up to a point: the experimental redistribution can only *ape* ('mimic' would be too much here) the redistribution of tax yields in the real world. Who gets what of the redistributed money in the experiment is totally transparent, for example, whereas tax money affects our lives in many indirect and hardly quantifiable ways. The trade-off between evading and paying taxes is easily computable in the experiment. The money is distributed to a small number of people, among which there probably at least a couple of friends, and so on. Importing features of the target here does not carry us a long way towards external validity because we *know* that the main dependent variable has been constructed 'artificially', and we are aware of its limitations with respect to the real thing.

The unexpected effect, in contrast, seems more promising. The idea is that *if* $X_1,\ldots,X_n$ are *really* the only variables that were artificially constructed by the experimenter, then the unexpected, residual effect is likely to be the consequence of some non-purely-experimental factor. An analogy here may help: cosmic microwave radiation was first observed in 1964 by Penzias and Wilson, two scientists at Bell Labs, while working on a problem of telecommunication technology. Perhaps the echo effect is like the isotropic radio background detected by Penzias and Wilson. The radiation is a left-over from the Big-Bang and fills the space everywhere in the universe. Regardless of where you are, it is there, although its properties may be in some circumstances be difficult to detect due to other disturbing factors and local circumstances. Such phenomena often emerge as residuals that cannot be imputed to the experimental procedures or other known factors, and prove to be extremely robust to measurement and experimental manipulation.

The analogy with the echo effect should be clear: first you observe something that you don't think has been created by the experimental procedures; then, by checking the robustness of the phenomenon to changes in other variables, you become more confident that the phenomenon is indeed a general feature of human psychology. The checking is important because the whole inference rests on a crucial background assumption: that no other 'artificial' factor besides $X_1,\ldots,X_n$ has been inadvertently built into the experiment. This assumption is credible if the experiment has been designed with enough care, and in part depends on the experience of the experimenter and her detailed knowledge of her system. But no matter how much experienced, some checking is necessary, and the scientific community will not be convinced until most attempts to 'make the effect go away' have failed.[23]

The generalisability of the echo effect to *specific* cases, nevertheless, remains an empirical conjecture, which has to be further validated by empirical investigation case-by-case. The experimental economist, in a case like this, proves the existence of a phenomenon, which is likely to be relevant to the policy-maker. The experimenter cannot guarantee that the phenomenon will be actually relevant in a specific case because the effect may be neutralised by some context-specific factor, but can signal a possibility. The actual effectiveness of the policy (repeated auditing, in this specific case) will depend on a number of features of the specific economic system at stake (the target system, in the terminology used in this paper).

[23] See also Galison (1987) for a similar form of reasoning in physics. Another intuitive analogy can be drawn with the way in which econometricians detect the existence of factors not explicitly modelled in the regression equations.

9. The library of phenomena

The examples we have described above are representative of a number of experiments (the majority, perhaps) performed in economics. In some happy cases the experimenter can go all the way from the model on the far left to the target system on the right of Figure 2. But these cases are quite rare. Most cases of experimentation involve inferences to generic circumstances rather than to specific situations. This is because the target is left unspecified, or cannot be studied properly for lack of data. Experimental Economics nevertheless helps the applied scientist by providing a '*library of phenomena*': a list of possible effects, biases, heuristics which can then be used in concrete applications. Each application then is a matter of examining the specific characteristics of the target domain, and based on this specific knowledge, evaluate the relevance of the phenomena found in the library, case by case.

This way of framing the problem recovers a basic distinction, between 'pure' and 'applied' science, while defending experimental economists from the charge of pursuing futile research. Although a research programme must eventually end in application, this need not be so for a single experiment. A single experiment may just highlight a phenomenon or cause-effect link, to be later exploited by applied scientists when they deal with specific cases. In any case, to export a phenomenon requires very detailed knowledge of the domain of application. Since the required knowledge is context-specific and probably generalisable only up to a point, it is reasonable to have a division of labour between applied scientist and experimenter.

Unlike some of its neighbour disciplines, like experimental psychology, that are widely used to resolve concrete problems (in the real world), the art of applying experimental economics is still underdeveloped.[24] Unlike psychologists, experimental economists grew within (and had to defend themselves from) a scientific paradigm that gives enormous importance to theory. This is probably why it was easiest and most effective from a rhetorical viewpoint to present experimental economics as primarily devoted to theory-testing. We have tried to show that this view is mistaken, and that the role of experimental economics is to mediate between abstract theory and problem-solving in the real world. In many respects experiments resemble models, for they are systems that are artificially isolated from the noise of the real world – but with the added bonus of a higher degree of concreteness. Like models, experimental results must eventually be applicable to real-world circumstances. As we have tried to show, the relevance of results may be indirect, and it is unreasonable to impose the requirement that the experimental validity of each single experiment be proven rigorously. In many cases, experimenters contribute to the 'library' of phenomena that the applied scientist will borrow and exploit opportunistically on a case-by-case basis.

References

Ankeny, Rachel (2001), "Model Organisms as Models: Understanding the 'Lingua Franca' of the Human Genome Project", *Philosophy of Science* 68 (Proceedings): S251-S261.

Binmore, K. (1999) "Why Experiment in Economics?", *Economic Journal,* 109, pp. F16-F24.

[24] For some examples from psychology, see e.g. Fischhoff (1996).

Brunsvik, E. (1939), “Probability as a determiner of rat behavior”, *Journal of Experimental Psychology*, 25, 175-197.
Brunsvik, E. and H. Herma (1951), “Probability learning of perceptual cues in the establishment of a weight illusion”, *Journal of Experimental Psychology*, 41, 281-290.
Cartwright, Nancy (1999), *The Dappled World*. Cambridge: Cambridge University Press.
Collins, H.M. (1985) *Changing Order,* Beverly Hills, Sage.
Davis, D.D. and C.H. Holt (1993) *Experimental Economics,* Princeton, Princeton University Press.
Fischhoff, B. (1996), “The real world: what good is it?”, *Organizational Behavior and Human Decision Processes* 65, pp. 232-248.
Fodor, J.A. (1974) “Special Sciences (Or: The Disunity of Science as a Working Hypothesis)”, *Synthese*, 28, pp. 97-115.
Frankfort-Nachmias, Chava and David Nachmias (1996), *Research Methods in the Social Sciences*. London: Arnold.
Friedman, M. (1953) “The Methodology of Positive Economics”, in *Essays in Positive Economics*, Chicago, University of Chicago Press.
Friedman, D. and S. Sunder (1994) *Experimental Methods: A Primer for Economists*, Cambridge, Cambridge University Press.
Galison, P. (1997) *Image and Logic*, Chicago, University of Chicago Press.
Giere, R.N. (1988) *Explaining Science*, Chicago, University of Chicago Press.
Gooding, David (1990), *Experiment and the Making of Meaning*. Dordrecht: Kluwer.
Guala, F. (1998) “Experiments as Mediators in the Non-Laboratory Sciences”, *Philosophica,* 62, pp. 901-918
Guala, F. (1999) “The Problem of External Validity (Or ‘Parallelism’) in Experimental Economics”, *Social Science Information*, 38, pp. 555-573.
Guala, Francesco (2001), “Building Economic Machines: the FCC Auctions”, *Studies in History and Philosophy of Science,* 32, pp.453-477.
Hempel, C.G. (1965) *Aspects of Scientific Explanation*, New York, Free Press.
Jarvik, M. E. (1951), “Probability learning and a negative recency effect in the serial anticipation of alternative symbols”, *Journal of Experimental Psychology*, 41, 291-297.
Kagel, John and Daniel Levin (1986), “The Winner’s Curse Phenomenon and Public Information in Common Value Auctions”, *American Economic Review* 76: 894-920.
LaFollette, Hugh and Niall Shanks (1995), “Two Models of Models in Biomedical Research”, *Philosophical Quarterly* 45: 141-160.
Mayo, Deborah (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Mittone, L. (2002) “Dynamic Behaviour in Tax Evasion”, CEEL Working Paper 3-02, University of Trento.
Morgan, Mary and Margaret Morrison (eds. 1999), *Models as Mediators*. Cambridge: Cambridge University Press.
Nagel, E. (1961) *The Structure of Science*, New York, Harcourt Brace and Wold.
Plott, C.R. (1991) “Will Economics Become an Experimental Science?” *Southern Economic Journal,* 57, pp. 901-919.
Plott, C.R. (1995) “Rational Individual Behaviour in Markets and Social Choice Processes: The Discovered Preference Hypothesis”, in K.J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds.) *The Rational Foundations of Economic Behaviour*, London, Macmillan.
Plott, C.R. (1997) “Laboratory Experimental Testbeds: Application to the PCS Auction”, *Journal of Economics and Management Strategy*, 6, pp. 605-638.
Popper, K.R. (1934/1959) *Logik der Forschung,* Vienna, Springer; Engl. transl. *Logic of Scientific Discovery*, London, Hutchinson.
Popper, K.R. (1957) “The Aim of Science”, *Ratio*, 1, pp. 24-35; reprinted in *Objective Knowledge*, Oxford, Clarendon Press, 1972.
Smith, V.L. (1982) “Microeconomic Systems as an Experimental Science”, *American Economic Review,* 72, pp. 923-955; reprinted in Smith (1991).
Sugden, R. (2000) “Credible Worlds: The Status of Theoretical Models in Economics”, *Journal of Economic Methodology,* 7, pp. 1-31.
Webley, P., Robben, H.S.J., Elffers, H., and Hessing, D.J. (1991). *Tax evasion: an experimental approach,* Cambridge: Cambridge University Press.
Wilde, L.L. (1981) “On the Use of Laboratory Experiments in Economics”, in J.C. Pitt (ed.) *Philosophy in Economics*, Dordrecht, Reidel.