Experiments in economics: External validity and the robustness of phenomena

Francesco Guala and Luigi Mittone

Abstract External validity is the problem of generalizing results from laboratory to non-laboratory conditions. In this paper we review various ways in which the problem can be tackled, depending on the kind of experiment one is doing. Using a concrete example, we highlight in particular the distinction between external validity and robustness, and point out that many experiments are not aimed at a well-specified real-world target but rather contribute to a ‘library of robust phenomena’, a body of experimental knowledge to be applied case by case.

Keywords: experiments, methodology, theory and evidence, external validity, pure and applied science.

JEL Codes: B41, C90

1 INTRODUCTION

Laboratory experiments have different purposes and motivations, and experimental economists have proposed various taxonomies to capture this diversity, e.g. Smith (1982, 1994), Friedman and Sunder (1994), Sugden (2005). Among these, the most popular one is probably Alvin Roth’s—sufficiently articulated to capture the variety of experimental practice, but simple enough to unify many experiments under a few encompassing categories. Roth (1986, 1988, 1995) proposes a three-fold classification of economic experiments based on their primary goals: (1) Speaking to theorists; (2) Searching for facts; and (3) Whispering in the ears of princes.

Experiments in the first category (‘Speaking to theorists’) test hypotheses that have been derived from formally specified models or theories. The second category, ‘Searching for facts’, includes experiments devoted to investigating phenomena or empirical regularities that cannot be explained by existing theoretical models. The third category, finally (‘Whispering in the ears of princes’), includes experiments devoted to illuminating or supporting policy-making. This paper discusses how such different types of experiments deal with the key problem of external validity of experimental results. External validity, in its original formulation, is a ‘question of generalizability: to what populations, settings, treatment variables, and
measurement variables can [an experimental effect] be generalized?' (Campbell and Stanley 1966: 5). 1

Historically, scepticism about external validity has constituted a major obstacle on the way towards the acceptance of experimental economics as a genuine scientific discipline. Together with impressive confirmations of the predictions of some economic models, experimental economics has produced stunning refutations of some basic theoretical principles. This is to be expected from a genuinely empirical discipline, but reactions to these findings have been varied and conflicting. Some economists have welcomed experimental anomalies as providing the ultimate proof that the fundamental principles of mainstream economic theory are flawed and need drastic revision. Others remain unconvinced and respond with 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.

Experimental economists are annoyed by such remarks, especially when they are formulated in the abstract (i.e. they do not specify what the difference between laboratory and real-world economies exactly amounts to), and are put forward by people who do not have a deep understanding of laboratory practice. Yet, there is something in the above critique which should be taken seriously. From a philosophical point of view, external validity is a puzzling issue which has been rarely discussed in depth in the methodological literature not only on economics, but on experimental science in general. Part of the reason is that experimenters themselves have argued occasionally that external validity is not a problem at all, or that it can be easily solved by means of simple methodological tricks. The arguments reviewed in the next section belong precisely to this category and are worth discussing in some detail.

2 ‘SPEAKING TO THEORISTS’ AND EXTERNAL VALIDITY

According to the ‘theory-testing view’ of science, as we shall call it, theorists propose, and experimenters control (and, sometimes, dispose). Some experimental economists (e.g. Plott 1991; Rubinstein 2001) speak as if all experiments were devoted to the investigation of scientific theories. According to Charles Plott, experimental economics’ revolutionary achievement consisted in shifting the focus from whether an experiment replicates a real-world system accurately to whether it accurately tests a theory or model. Experimental economists have thus removed two ‘constraints’ that stood in the way of laboratory research:

The first was a belief that the only relevant economies to study are those in the wild. The belief suggested that the only effective way to create an experiment would be to mirror in every detail, to simulate, so to speak,
some ongoing natural process. ... As a result the experiments tended to be dismissed either because as simulations the experiments were incomplete or because as experiments they were so complicated that tests of models were unconvincing. ... Once models, as opposed to economies, became the focus of research the simplicity of an experiment and perhaps even the absence of features of more complicated economies became an asset. The 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. (Plott 1991: 906)

This view should not be taken as representative of experimental economists’ position in general. Other economists (like Alvin Roth, as we have seen) allow for more diversity in experimental practice. Although we tend to side with the pluralists, we find it useful here to start from the theory-testing view, if only for expository purposes. The questions we would like to discuss are very basic: what sort of conclusions can be derived from the test of an economic model in the laboratory? In particular, can experiments teach any general lesson, applicable to non-experimental circumstances too?

Let us start from an admittedly oversimplified version of theory-testing: let us assume that science is a game with two players (see Figure 1). On the one hand we have theory (which in economics usually takes the form of a formal model), on the other there is 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.

However, this view is absurd. In particular, it presupposes that everything that takes place on the right-hand side is relevant for the appraisal of economic models. In fact 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 its intended domain of application.3 There are two standard replies to this challenge, which, unfortunately, are both unconvincing.

(a) Economic theories are general in scope of application:
This claim has been defended, among others, by Charles Plott:

The logic is as follows. General theories must apply to simple special cases. The laboratory technology can be used to create simple (but real) economies. These simple economies can then be used to test and

Figure 1  Simplified theory-testing view
evaluate the predictive capability of the general theories when they are applied to the special cases. In this way, a joining of the general theories with data is accomplished. (Plott 1991: 902)

A staggeringly large number of theories exist. One purpose of the laboratory is to reduce the number by determining which do not work in the simple cases. The purpose is also to improve the models by exploring how a model might be changed to make it work better in the simple cases. General models, such as those applied to the very complicated economies found in the wild, must apply to simple special cases. Models that do not apply to the simple special cases are not general and thus cannot be viewed as such. (Plott 1991: 905)

So, 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. Imposing requirements of generality and universality on ‘genuine’ scientific theories is common practice in physics-oriented philosophy of science. However, it is doubtful whether the requirement that theories should be general in scope of application should be extended to other scientific disciplines too. In fact, most economic models describe mechanisms and phenomena embedded in fairly specific institutional settings. It is fairly 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, but are not instantiated 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 when the background conditions are ‘right’, in transparent settings with learning and repetition, for instance, but can fail in less than ideal environments (as recognized, curiously, by Plott (1996) himself; see also Binmore (1999)).

Of course, one may reformulate the argument in normative form: perhaps economic models are not general in scope and application, but they should be. In other words, the generality requirement, would be a desideratum, an ideal that guides the development of science. However, there are good reasons to believe that the ideal may be impossible to satisfy in disciplines like economics. Unlike physics, which is concerned with the discovery of the most basic properties of matter, sciences such as 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 theoretical principles 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. It would probably require a massive injection of psychology and neurophysiology in order to capture the most fundamental mechanisms that underlie human judgement and decision-making; and probably also a tighter partnership with politics, morality and the study of culture in general, in order to understand the social factors that influence the interaction of individuals in normatively charged circumstances. At any rate, the reduction to neuroscience is far beyond our present capacities, and the partnership with other neighbouring disciplines underdeveloped. We need a science of economic phenomena in the meantime, and to set unrealistic goals may do more harm than good by taking economists’ attention away from what can be really achieved.8

(b) **Economic experiments should mirror the assumptions of models:** According to the second argument if the theory is simple, the experiment is allowed to reproduce its initial conditions (or assumptions) in the laboratory. If the model is too simple to predict the observations, this is a problem for the theorist, not for the experimenter:

If [an experiment’s] purpose is to test a theory, then it is legitimate to ask whether the elements of alleged ‘unrealism’ in the experiment are parameters in the theory. If they are not parameters of the theory, then the criticism of ‘unrealism’ applies equally to the theory and the experiment. If there are field data to support the criticism, then of course it is important to [modify] the theory to include the phenomena in question, and this will affect the design of the relevant experiments. (Smith 1982: 268)

These claims shift the burden of proof onto the critics of experimental economics. They can be paraphrased into a powerful and indeed very reasonable argument: you say that experiments have external validity problems; but then consider that theoretical modelling (the most prestigious activity in economics, where most Nobel Prizes are won) faces exactly the same challenge. So if you think that theory is valuable, experiments must be valuable too. But the argument does not show that we should not worry about external validity. Rather, it points out that external validity is an important issue for experimenters and theorists alike, that theorists should worry about it too. Knowing that experimental economists are not playing a solipsistic game of their own is hardly a relief, if the alternative is playing a game with theorists only.
Moreover, an important implication of the argument outlined in the previous 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. 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.

Indeed, this 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 reproduction of a phenomenon (see 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. To summarize, according to these considerations, to shift the focus from real economies to theoretical models hides significant questions about the significance of experimental results. 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 cast the role played by experiments and models in a different light.

3 EXPERIMENTS AS MEDIATORS

When we fund medical research, for example, on the effects of a new drug, we expect eventually to obtain results that are 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 their target (the real-world systems in model’s ‘intended domain’ of application). This picture of science is represented in Figure 2.

According to such a view, experimental systems are just an intermediate step in the route from pure theory to their target economic phenomena. They are ‘mediators’, in the sense that they help bridge the gap between models and their intended domain of application. The worst aspect of the theory-testing view is that it induces us to think of theoretical models and experimental systems as radically different entities, where in reality they are not. Both models and experimental systems should be thought of as structures that we are studying and using to understand reality. Models are sets of (abstract or, less frequently, concrete) entities endowed with
properties and relations between them. Experimental systems are obviously more concrete than models, and ideally should be closer to the intended domain of application: they should share more properties (including materiality) with the systems we are eventually interested in understanding (the target economies). But they are not necessarily the target system, and to move from experiment to target usually requires an inference of some sort. Just like the hypothesis that features of the experimental system are adequately represented in a given model, the hypothesis that features of the target domain are adequately represented in an experimental system is fallible. This hypothesis must be tested empirically, and the evidence we have got is usually insufficient to logically entail the claim that the experimental system is a valid representative (in some scientifically relevant respect) of the target system. The inference from the data to the claim is inductive or ampliative—experimenters can be (and occasionally are) wrong in making validity claims.

There is quite a lot of tacit and explicit knowledge about how to bridge the gap between models and experimental systems, 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 testing various causal hypotheses in isolation. 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, the influence of background factors etc. have been checked one by one.

At a more concrete level, the task of establishing the internal validity of an experimental result depends on a much context-specific knowledge and techniques. 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. Frankfort-Nachmias and Nachmias 1996) includes chapters on how to control preferences, how to rule out undesired effects, and eventually how to draw inferences from observed statistical frequencies to causal relations between properties of the experimental system. Of course, much depends on the intuition and the creativity of the experimenter, but

---

**Figure 2** Experiments as mediators

- Model
- Experimental System
- Target
- Real World

---

*External validity and robustness* 501
the basic strategies are well known. In contrast, we find very little explicit advice about how to bridge the second gap, the one from experimental to target phenomena (external validity problem).

4 ‘WHISPERING IN THE EARS OF PRINCES’ AND EXTERNAL VALIDITY

There has been a lot of excitement recently about the applicability of experimental economics to solve complicated problems of market regulation. The most famous examples are the experiments that informed the design of the auctions for mobile phone licences, first devised by the Federal Communication Commission (FCC) in the US and then replicated in several other countries. These are paradigmatic examples of type-3 experiments, devoted to policy-making or, in Roth’s terminology, ‘Whispering in the ears of princes’. Since the story of the FCC experiments has already been told elsewhere, we shall not repeat it here (cf. Plott 1997; Guala 2001, 2005; Klemperer 2004). For our present concerns it should be noted that the construction of these auctions was an example of direct attack on the problem of external validity. The target itself (the market for portable communication systems) was shaped in order to fit the experimental prototypes in the laboratory. Again, there is nothing peculiar about this, it is just what happens with most technology, from space probes to the TV sets in our homes.15

But economists usually are not so lucky: they can rarely shape the world as they wish. It is more common for them to follow the opposite strategy. Instead of shaping the world on the experimental system, the latter may be designed to resemble the target 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. In other words, we need to identify 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 selling leases to drill oil in the Outer Continental Shelf (Gulf of Mexico) (cf. Kagel and Levin 1986; Guala 1998).

This strategy tackles the problem of external validity effectively. In fact if you observe phenomenon X in the laboratory, but you are not sure about its generalizability to generic real-world circumstances, it is difficult to tackle the problem constructively. You can do much better in contrast 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 relevant (causal) dissimilarity between the experimental and the target system. Thus, the obvious way to proceed is to modify the experimental system 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 in some important respect from a descending auction, use the former in your experiment; and so on.

5 ‘SEARCHING FOR FACTS’ AND EXTERNAL VALIDITY

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 (for example, think of 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 many other cases of experimentation this can be done only to a much lesser extent. Moreover, the idea that external validity inferences are always generalizations aimed at specific targets (‘from the particular to the particular’, to use a traditional philosophical formula, a view defended, for example, by Guala 2003) seems to leave out experiments that are not performed with a specific target in mind. Yet, 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, some experiments concerning tax evasion conducted by one of us at the University of Trento (cf. Mittone 2002).

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 (e.g. Webley et al. 1991: 39–47). 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 experimental task, rather than take it seriously), of absence of social incentives (your family and friends do not know, and do not care, whether you are busted by the experimental taxman), absence of social actors (no lawyers, no accountants), and so on. Notice that 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 hence 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. However, the theory-testing view is too narrow, and this case just provides more evidence that it is. Following Roth, these are more adequately characterized as type-2 experiments, aimed at ‘searching for facts’.

In Mittone’s experiments subjects were provided with a set of parameters (income, the current income tax rate, the probability of being audited)
and were asked to decide how much tax to pay in each round. The experiment was concerned with dynamic choice, and thus was extended over a series of sixty 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 analogous to free riding. Given that (contrary to standard theory) cooperative behaviour has been extensively observed in public goods experiments, the redistribution of yield may have the effect of reducing the rate of evasion behaviour. In fact this phenomenon was observed in the experiment.16

However, this result is just the beginning of the story: as we said, theory-testing is not the only, nor the main goal of experimental economics. A most 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. Phenomena are empirical regularities that may or may not be adequately explained by the available theories. Although ideally all phenomena should eventually be explained theoretically, they can be ‘free standing’ for a long time, detached from any theory or even informal explanation of their occurrence. Sometimes phenomena become somehow uninteresting for the scientific community and ‘die’ unexplained: researchers simply stop searching for an explanation. However, a free-standing phenomenon is always attached to its design: a description of the kind of circumstances in which it is likely to be produced and observed. In a recent paper, Robert Sugden (2005) introduces the term ‘exhibit’ to denote the phenomenon-design pair. Borrowing an expression first used by Daniel Kahneman, he says that an exhibit is a ‘bottled phenomenon’, a particularly appropriate metaphor, for it reminds that phenomena always come with their ‘bottle’: a (sometimes rough but always informative) description of the experimental design.

Although phenomena do not have to be strictly speaking experimental, the latter are particularly interesting. Unlike field observation, laboratory investigation 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) it is robust to changes in background factors (or a demonstration of what factors the phenomenon depends upon). 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 two phenomena that emerged from the tax experiments: we shall call them the ‘bomb crater’ and the ‘echo’ effect.
6 BOMB CRATERS AND ECHO

They say that troops under heavy enemy fire hide in the craters of recent explosions, for they believe it is highly unlikely that two bombs will fall exactly in the same spot within a short time period. 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.

The experimental task is simple: subjects can decide how much tax they want to pay, where the tax due is determined as a proportion of an exogenously fixed income. The tax due varies during the experiment (see the ‘stepwise’ solid line in Figure 2). After she has made her decision, a lottery determines randomly whether a subject will be audited or not, and the subject is aware of the probability of being audited. As can be seen from Figure 2, evasion increases after each audit (the dotted line, average tax paid, drops after each column).

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 occur in a number of other situations that have nothing to do with taxes and audits. Figure 4 represents

![Graph](image)

**Figure 3** The ‘bomb crater’ effect
a totally abstract game that has been tried repeatedly in the University of Trento’s experimental lab. 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 equal to 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.17

Now, it is possible to use this completely abstract, decontextualized 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.

What does this tell us about the generality of the bomb crater effect? The most plausible answer is that we are dealing with a robust bias, which tends to arise whenever subjects have to do with probabilistic reasoning of this kind.18 To establish robustness is to establish a sort of generality, to a set of

![Figure 4 The ‘wheel of numbers’](image)
situations that are somehow 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 provides no precise indication of the situations in which it will occur and those in which it will not. Stylized 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 further 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 taxpayers more often than the others. This strategy (if known to taxpayers, which is an empirical hypothesis of course) may be enough to attenuate or even eliminate 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 5) or the late rounds (Figure 6) 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,

![Figure 5](image-url)  
*The effect of early punishment*
and apparently it takes time for them to revise their attitude. Again, this sort of phenomenon is robust to changes in design conditions. Moreover, we can think of several examples from real life that seem to support the generalizability of the ‘punishment’ effect. In Italy, for example, where police officers are well known for their inconsistency in fining traffic offenders, car drivers seem to proceed at lower speed on roads upon which the police have 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.

7 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 \((X_1, X_2, \ldots, X_n)\) 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 simple 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 factors 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 make some. The idea of having yield redistribution can also be seen as an attempt to ‘import’ into the experiment some real-world features and thus increase the external validity of the experimental results (along the lines sketched in section 5 above). But this attempt is convincing only up to a point: the experimental redistribution can only ape (‘represent’ would be too strong a word 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 whom there are 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 or manipulated by the experimenter, then the unexpected, residual effect is likely to be the consequence of some non-purely-experimental factor. A physical analogy here may be of some help: cosmic microwave radiation was first observed in 1964 by Arno Penzias and Robert 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 leftover from the Big-Bang and fills the space everywhere in the universe. Regardless of where you are, it is there, although its properties may in some circumstances be difficult to detect due to disturbing factors and other local circumstances. Such phenomena often emerge as residuals that cannot be imputed to the experimental procedures or other known factors, and are often extremely robust to measurement and experimental manipulation.

The analogy with the echo effect should be clear: first you observe something that you do not 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 robust 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.20

The generalizability of the echo effect to specific cases, nevertheless,
remains an open question, which has to be further validated by empirical
investigation on a case-by-case basis. The experimental economist, in a case
like this, establishes 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 neutralized by some context-specific factor; but still, she 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).

8 THE LIBRARY OF PHENOMENA

The problem of generalizing from the laboratory then takes different forms
in experiments of different kind. 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. These ‘robustness tests’ nevertheless help the applied
scientist by contributing to a ‘library of phenomena’: a list of possible
effects, biases, heuristics that 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 theoretical models and of the phenomena found in the
library, case by case.

This way of framing the problem has the advantage of recovering a basic
distinction, between ‘pure’ and ‘applied’ science, while defending experi-
mental economists from the charge of pursuing futile research. Although
ideally a research programme should eventually lead to 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 in the real world requires very detailed knowledge of the
domain of application. Since the required knowledge is context-specific and
probably generalizable only up to a point, it is reasonable to have a division
of labour between ‘applied’ and ‘pure’ experimenter.
Unlike some of its neighbour disciplines, such as experimental psychology, which are already widely used to address concrete problems (in the real world), the art of applying experimental economics is still underdeveloped. Unlike psychologists, experimental economists work within (and have to defend themselves from) a scientific paradigm that gives enormous importance to theory. This is probably why it was easier and more effective from a rhetorical viewpoint to present experimental economics as primarily devoted to theory-testing. We agree with Alvin Roth that this view is too narrow. We have argued that often the role of experimental economics is to mediate between abstract theory and problem-solving in the real world. In many respects experimental systems resemble models: 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 systems must eventually bear some relation to real-world circumstances. However, as we have tried to show, the relevance of experimental results may be indirect, and it is unreasonable to impose the requirement that the external validity of each single experiment be demonstrated with reference to a specific kind of situation. In many cases experimenters contribute to the ‘library’ of phenomena that the applied scientist will borrow and exploit on a case-by-case basis.

Francesco Guala and Luigi Mittone
University of Exeter
f.guala@ex.ac.uk
University of Trento
luigi.mittone@economia.unitn.it

ACKNOWLEDGMENT
Parts of this paper derive from a seminar Francesco gave at the University of East Anglia in October 2001. The authors wish to thank Riccardo Boero, John Broome, Dan Hausman, Uskali Mäki, audiences at East Anglia and at the 2002 Experimental Economics Workshop in Siena for their insightful comments and critiques. Francesco Guala acknowledges financial and logistic support from the Cognitive Science Laboratory at the University of Trento. The experiments were run at CEEL with the invaluable help of Marco Tecilla. The usual caveats apply.

NOTES
1 When this problem has been discussed by experimental economists, it has sometimes been renamed the problem of ‘parallelism’, e.g. Smith (1982), Wilde (1981).
Indeed, some of Plott’s own work seems to fall squarely under the ‘whispering in the ears of princes’ category. However, his position is that even experiments aimed at informing policy-making involve the testing of models. Elsewhere (e.g. Plott 1981: 134) he suggests that economic engineering is best seen as the problem of testing a theoretical mechanism and turning it into an operational institution.

On the difficulty of specifying exactly the intended domain of application of an economic model, and the problems this causes to experimental testing, see Cubitt (2005).

There are some dissenters even regarding physics, e.g. Cartwright (1999).

In fact, part of the contribution of experimental economics has been to identify some of the conditions that make the instantiation of such phenomena and the working of such mechanisms possible.

This obviously suggests that rational behaviour might not take place in several ordinary economic contexts, something that even some rational choice theorists are beginning to recognize (cf. Binmore 1999).

This position is strongly defended, for example, by Popper (1957).

It should be noted 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 outlined in the main text see Fodor (1974).

This is, incidentally, one rationale behind Milton Friedman’s (1953) famous thesis on the inevitable ‘unrealisticness’ of the assumptions of economic models.

‘Applicability’ is a complex concept and will be discussed more carefully in section 8; for the time being we ask the reader to take it unproblematically and intuitively.

See also Guala (1998, 1999, 2005). The term ‘mediators’ has been borrowed from Morgan and Morrison (1999). The theory of scientific models outlined below is close to the so-called Semantic View of theories, e.g. van Fraassen (1980). For a similar view of modelling in economics, see Sugden (2000).

Mäki (2005) defends the stronger thesis that ‘experiments are models and models are experiments’. In contrast, we think that there are some obvious important disanalogies, in particular that experimental systems should in some crucial respects be made ‘of the same stuff’ as the target systems. For a more detailed discussion of this ‘materiality’ requirement, see Guala (2002).

We say ‘not necessarily’ because if economic principles are interpreted as general and universal laws then what happens in an experiment clearly belongs to the intended domain of application. We choose this formulation for the sake of generality, even though we have argued against this interpretation in Section 2 above.

Guala (2005: chs 3–6) includes a more detailed discussion of the methods used to solve internal validity problems.

Bruno Latour (1984) and Nancy Cartwright (1999) are well known in the science studies literature for their emphasis on the fact that the most reliable applications of science are obtained by shaping the real world so as to resemble laboratory conditions.

In addition, 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. See Mittone (2002) for more details.
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.

This bias is related to (perhaps is 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 1930s (cf. Brunsvik 1939; Brunsvik and Herma 1951; Jarvik 1951).

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

See also Galison (1997) 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. For an analysis of this sort of reasoning in terms of ‘eliminative induction’ see Guala (2005).

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

REFERENCES


