Economics and Philosophy, **16** (2000), 47–75

ARTEFACTS IN EXPERIMENTAL ECONOMICS: PREFERENCE REVERSALS AND THE BECKER–DEGROOT–MARSCHAK MECHANISM

Francesco Guala
University of Exeter

1. INTRODUCTION

Controversies in economics often fizzle out unresolved. One reason is that, despite their professed empiricism, economists find it hard to agree on the interpretation of the relevant empirical evidence.[1] In this paper I will present an example of a controversial issue first raised and then solved by recourse to laboratory experimentation. A major theme of this paper, then, concerns the methodological advantages of controlled experiments. The second theme is the nature of experimental artefacts and of the methods devised to detect them. Recent studies of experimental science have stressed that experimenters are often merely concerned about determining whether a certain phenomeonon exists or not, or whether, when, and where it can be produced, without necessarily engaging in proving or disproving any theoretical explana-

I would like to thank Dan Hausman, Philippe Mongin, Matteo Motterlini, John Worrall and three referees for their comments on previous drafts of this paper. This research has been supported by a Marie Curie fellowship, a British Academy award, the Centre for Philosophy of Natural and Social Science at the London School of Economics, and the centre THEMA at the University of Cergy-Pontoise.

[1] I am not suggesting that this is the *only* reason; for more views on this topic, see for instance the symposium in the *Journal of Economic Methodology*, Vol. 1, June 1994. Hausman and Mongin (1998) discuss the question using preference reversals as one of their case studies.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

tion of the phenomenon itself.[2] In this paper I shall be concerned mainly with such a case, and focus on the example of preference reversals, a phenomenon whose existence was until quite recently denied by the majority of economists. Their favourite strategy consisted in trying to explain the phenomenon away as an artefact of the experimental techniques used to observe it. By controlled experimentation, as we shall see, such an interpretation has been discredited, and now preference reversals are generally accepted as real. The problem of distinguishing an artefact from a real phenomenon is related to methodological issues traditionally discussed by philosophers of science, such as the theory-ladenness of observation and Duhem's problem. Part of this paper is devoted to clarifying these two philosophical problems, and to arguing that only the latter is relevant to the case in hand. The solutions to Duhem's problem devised by economic experimentalists will be presented and discussed. I shall show that they belong in two broad categories: independent tests of new predictions derived from the competing hypotheses at stake, and 'no-miracle arguments' from different experimental techniques delivering converging results despite their being theoretically independent.

2. PREFERENCE REVERSALS

Preference reversals (PR from now on) were first observed by two psychologists at the Oregon Research Institute, Paul Slovic and Sarah Lichtenstein. They conjectured that, far from constituting the stable substratum from which all economic behaviour arises, preferences are 'constructed' and vary from context to context. In a PR experiment, subjects are asked in separate tasks to choose among two bets and to price them. The pairs of lotteries have a common feature: they all consist of a bet with a high probability of winning a moderate amount of money and a low probability of losing a small amount (called the 'P-bet': 0.9 to win $5, 0.1 to lose $1, for example); and a bet with a low probability of winning a larger sum and a high probability of losing a small sum (the '$-bet': e.g., 0.3 to win $15 and 0.7 to lose $2). Moreover, they have approximately the same expected monetary value. Slovic and Lichtenstein conjectured that 'bidding and choice involve two quite different processes that involve more than just underlying utilities of the gambles' (1971, p. 47).

On the basis of earlier experimental results, Slovic and Lichtenstein (1968), believed the act of choosing to be correlated with the probabilities, and the act of pricing with the monetary prizes of the gambles. An experiment was conceived explicitly to test the prediction that subjects

[2] Cf. for example Hacking (1983), Collins (1985), Franklin (1986), Galison (1987).

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

choose the P-bet but bid more for the $-bet. A pair for which this happens is said to exhibit a *reversal*. As a matter of fact, such patterns were observed (Lichtenstein and Slovic, 1971, 1973). The standard rate of reversals observed by Lichtenstein and Slovic, and then in later PR experiments, was between 70 and 80%. Not all reversals were of the kind predicted by Lichtenstein and Slovic, though: in a 'standard' PR experiment, from 15 to 25% of reversals are of the non-predicted (or 'asymmetric') type.[3]

Lichtenstein and Slovic (1971) performed three experiments. In order to control for possible disturbances due to lack of incentives, they used, in two of their experiments, an elicitation procedure known since the mid-sixties as the Becker–DeGroot–Marschak (BDM) mechanism. The BDM procedure is a tool devised to elicit the selling price of any kind of commodity, and as such has often been used to control subjects' preferences over lotteries. To elicit the certainty-equivalent of a lottery, in fact, a pay-off mechanism must be used to make sure that the price reflects the subject's *real* preference.

In a BDM elicitation, a subject is asked to state her reservation price, s, for a lottery with monetary payoffs (say, $[x, p; y, (1-p)]$); then, the lottery is auctioned, and if a buyer willing to bid a sum $b \geq s$ is found, the subject receives b; otherwise, the lottery is played out, and the subject receives a sum x or y according to the outcome. The experimenters draw the bidding sum b from a uniform distribution over some relevant set. It is easy to show that an expected utility maximizer must state his true selling price.[4]

The BDM mechanism is often used in conjunction with the so-called Random Lottery Selection (RLS) procedure. In general, experimental subjects are asked to perform more than just one task; instead of receiving an aggregate payment, each subject is rewarded according to the results of only one task selected at random at the end of the session. This procedure controls for 'endowment effects' (when a subject performs several tasks in succession, her preferences may vary because of changes in her wealth) and reduces experimental costs at the same time. In the case in hand, if the selected task is a choice one, a lottery will be simply played out; if it is a pricing task, the BDM mechanism is used.

When other experimentalists began to replicate Lichtenstein's and Slovic's findings, they also used the BDM and RLS procedures. David Grether's and Charles Plott's (1979) research, for example, was driven by the suspicion that PR may have been the product of some undetected

[3] For a non-technical presentation of the early research on the PR phenomenon, cf. Thaler and Tversky (1990). Hausman (1992) has used the debate on PR as an example of economists' dogmatic attitude towards disturbing empirical results; Tammi (1999) has analysed this controversy as a process of negotiation.

[4] The original proof is in Becker, DeGroot and Marschak (1964).

experimental effect. Despite great care in designing the experiment to control for all sorts of disturbances, however, they observed the same results Lichtenstein and Slovic had produced a few years earlier.

The historical and methodological significance of Grether and Plott's experiment has been discussed in depth elsewhere (Hausman, 1992). In this paper I shall focus on a later debate about the appropriate use of elicitation mechanisms and the observation of PR. Some of the best experimenters in economics devoted lots of time and effort to test the functioning of the BDM and RLS mechanisms – despite the fact that, as already mentioned, PR had been observed *with and without* these elicitation procedures. I shall for the time being bracket the question of why it took so long to convince economists that elicitation mechanisms were not to blame for the PR results,[5] and focus on the way experiments were used to shift the burden of proof towards the sceptics. The controversy on the BDM and RLS procedures is a paradigmatic example of disagreement concerning the interpretation of scientific evidence, and of how experimental ingenuity can eventually resolve it.

3. DATA, PHENOMENA AND ARTEFACTS

Bogen and Woodward (1988) have forcefully argued that scientists customarily explain *phenomena*, rather than data. Phenomena can be thought of as similar to what neopositivists called 'experimental laws',[6] regularities occurring in some specific experimental situation. Like many experimental laws, Bogen's and Woodward's phenomena are not directly observable. They are rather inferred from data.

> Data, which play the role of evidence for the existence of phenomena, for the most part can be straightforwardly observed. However, data typically cannot be predicted or systematically explained by theory. By contrast, well-developed scientific theories do predict and explain facts about phenomena. Phenomena are detected through the use of data, but in most cases are not observable in any interesting sense of the term. (Bogen and Woodward, 1988, pp. 305–6)

The PR phenomenon is an example of a 'phenomenon', or perhaps a class of phenomena, in Bogen's and Woodward's sense. To begin with, PR are not directly observable. We rather observe patterns of behaviour that appear prima facie incompatible with the claim that 'there exists a transitive scale of preferences underlying subjects' choices'. The data obtained in a typical 'PR experiment' may be represented as sentences like 'subject x has chosen the P-bet over the $-bet and priced the $-bet

[5] But see footnote 22 below.
[6] Cf. Nagel (1961, Ch. 5).

higher than the P-bet'. In order to obtain the PR phenomenon, one needs to assume, to begin with, that pricing and choosing convey genuine information about unchanging preferences – if the latter exist at all.

At this stage psychologists and economists part company. Economists, in fact, typically assume that the same preference structure underlies both pricing and choosing, whereas psychologists – as already mentioned – doubt that the idea of a stable preference scale is useful at all.[7] Following the 'economic' approach one is led to infer the existence of a genuinely intransitive preference structure. The phenomenon can then be represented as follows:

(PR) $P >_c \$ >_p P$, and $>_c = >_p$.

Some theorizing has thus taken place on the way from the observation reports to the phenomenon in the form above. Here, '$>_c$' stands for 'preference as emerging from choice', and similarly '$>_p$' for 'preference emerging from pricing'.[8] The first inference involves some assumptions about the correct functioning of our instruments of elicitation, whereas the latter involves a commitment to the principle of *procedure invariance* – the idea that all economically relevant behaviour is determined by the same preference scale, and thus that all economic behaviour can be used as evidence for inferring the structure of preferences. Theoretical and non-theoretical assumptions of this kind sanction the step from reports like 'subject x has chosen so-and-so' while 'subject y has priced so-and-so' to claims about preferences; or from observed apparent 'price-choice reversals' to real PR.

The data-phenomena distinction can help us to define the concept of '*artefact*' as it will be used in the next pages. In this sense (which, as we shall see in Section 10, is not the only one in scientific discourse) artefacts are interpretations, and mistaken ones, of a certain set of data. They are cases of misleading connections between data and phenomena, typically due to the method of observation: when the data are contaminated by some unknown factors, for instance; or when the scientist holds an incorrect theory of the experimental apparatus he is using.

[7] For reasons of simplicity, in this paper I shall mainly focus on the 'economic' interpretations of the PR experiments, thus disregarding those which question the very existence of a preference structure. I shall therefore often apply the PR label to intransitive patterns of preferences rather than (more correctly and more in general) to the class of all regularities compatible with the non-existence of a transitive preference scale. For some attempts to discriminate in the laboratory between the 'economic' and the 'psychological' interpretations, see Loomes, Starmer and Sugden (1989) and Tversky, Slovic and Kahnemann (1990).

[8] More precisely, $>_c$ and $>_p$ are binary relations on a lottery space, which satisfy at least an asymmetry condition.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

4. EXPLAINING PREFERENCE REVERSALS AWAY[9]

In the mid-eighties some economists began to argue that the PR phenomenon could have been an artefact of the experimental procedures used in order to control subjects' preferences: in other words, that inferences from data to $P >_p \$$ and $\$ >_c P$ were faulty. Charles Holt, Uzi Segal, and Edi Karni and Zvi Safra, independently and almost simultaneously, began to investigate theoretically the robustness of experimental procedures to violations of the axioms of expected utility theory. They pointed out that the controls used by Grether and Plott and other experimenters[10] 'are appropriate if the axioms of von Neumann–Morgenstern utility theory are satisfied' (Holt, 1986, p. 509). Their arguments were directed towards either the BDM procedure or the RLS procedure. The dependence of elicitation procedures on expected utility theory was hardly a new discovery: the inventors of the BDM mechanism knew and wrote explicitly that 'the procedure is based upon the [...] well-known ''expected utility hypothesis''' (Becker, DeGroot and Marschak, 1964, p. 226).

As we have seen above, in a PR experiment the subjects are asked to perform a number of tasks; from these, one is selected at random (RLS procedure); then, if necessary, the BDM mechanism is used. Using the probability calculus, a lottery in multiple stages can be reduced to a single-stage one, and expected utility theory requires that people's preferences in the multi-stage lottery be consistent with those in the reduced one. Formally, the *reduction principle* states that subjects are indifferent between a compound lottery $A = (X_1,q_1;\ldots;X_m,q_m)$, giving a chance q_i to participate in a lottery $X_i = (x^i_1,p^i_1;\ldots;x^i_{ni},p^i_{ni})$, and the reduced lottery

$$R(A) = (x^1_1,q_1p^1_1;\ldots;x^1_{n_1},q_1p^1_{n_1};\ldots;x^m_1,q_mp^m_1;\ldots;x^m_{n_m},q_mp^m_{n_m}).$$

Holt conjectured that subjects saw the PR experiment as a two-stage lottery of this sort: in the first stage, the task to be played out is randomly selected; in the event of pricing, there is a second stage, that is, the task is played out *via* the BDM mechanism. Then, Holt claimed that *if* they apply reduction but not the independence principle, some subjects who *prefer* the \$-bet to the P-bet may express *choices* that do not reflect their preferences. Behaviour observed via the RLS mechanism, then, may not reveal true preferences.

The suggestion of blaming violations of independence was in line

[9] Although I have tried to simplify the issue as much as possible, the contents of this section remain quite technical; it should be possible however to follow the general line of argument while skipping some of the details. The reader interested in the mathematical proofs should refer to the original papers cited below.

[10] Cf. Pommerehne, Schneider and Zweifel (1982), and Reilly (1982).

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

with the arguments put forward in that period by theorists working on other anomalies. Mark Machina's (1982) 'Generalized Expected Utility Analysis' was at the time a fully developed alternative approach to decisions under risk, which relaxed the principle of independence and allowed utility functions to be merely differentiable rather than strictly linear in the probabilities. Other approaches, like Chew and MacCrimmon's (1979) 'Alpha Utility' theory, or Quiggin's (1982) and Yaari's (1987) 'Expected Utility theory with Rank-Dependent Probabilities' (EURDP) were being developed, which similarly made do without independence or related principles.

Loosely, the *principle of independence* says that only the outcomes that distinguish two lotteries are relevant to the decision to be taken. More precisely, it says that if a lottery X is preferred to another lottery Y, then the compound lottery $(X, p; Z, 1-p)$ is preferred to $(Y, p; Z, 1-p)$. Violations of this principle were the first to be discussed by decision theorists, thanks mainly to Maurice Allais's (1953/1979) early experiments. It was therefore tempting to argue that the illusion of PR resulted from violations of the Allais kind. According to Holt, and more specifically Karni and Safra (1987), the BDM mechanism may be perceived by subjects as a two-stage lottery giving, among its outcomes, the possibility of playing out the priced gamble. Suppose the latter is $X = (4, 35/36; -1, 1/36)$ – one of the P-bets used by Grether and Plott (1979), typically subject to PR. If, by assumption, both the $\pi(X)$ – the real selling price – and b – the bidding price – are restricted to the 1000 equidistant values from 0 to 9.99, the following two-stage lottery results from the BDM procedure:[11]

$$A = \left(\left(4, \frac{35}{36}; -1, \frac{1}{36}\right), \frac{\pi(\mathrm{X})}{10}; \delta_{\pi(\mathrm{X})}, \frac{1}{1000}; \delta_{\pi(\mathrm{X})+0.01}, \frac{1}{1000}; \ldots; \delta_{9.99}, \frac{1}{1000} \right)$$

where the δ_i stand for degenerate lotteries with probability 1 of getting i, and $\pi(X)/10$ is equal to the probability of participating in X according to the BDM mechanism. The lottery A is equivalent to the tree in Figure 1.

By definition of a certainty equivalent (CE), we know that $\mathrm{X} \sim \delta_{CE(\mathrm{X})}$. Thus, by applying independence, there follows that

$$A \sim A' = \left(CE\left(4, \frac{35}{36}; -1\frac{1}{36}\right), \frac{\pi(X)}{10}; \delta_{\pi(X)}, \frac{1}{1000}; \delta_{\pi(X)+0.01}, \frac{1}{1000}; \ldots; \delta_{9.99}, \frac{1}{1000} \right).$$

The indifference above implies that agents see Tree 1 as equivalent to the tree in Figure 2.

The task faced by an agent participating in a BDM experiment, then,

[11] I follow here the presentation given by Keller, Segal and Wang (1993).

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

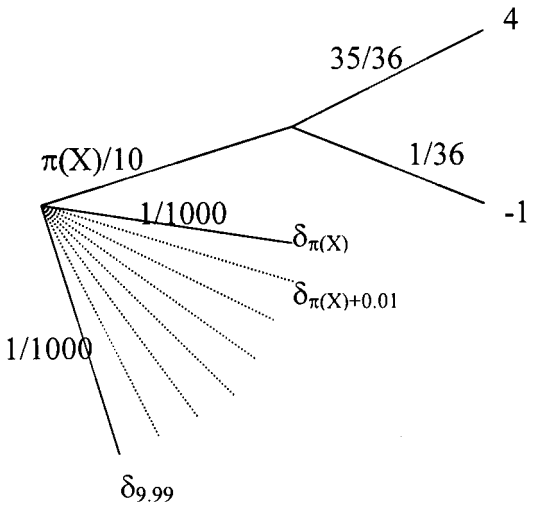

FIGURE 1. The lottery A

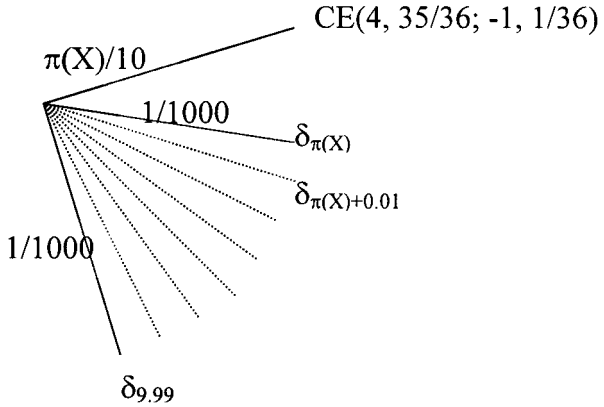

FIGURE 2. The lottery A′

is representable as a maximization problem: what is the value of $\pi(X)$ that maximizes the value of the lottery A'? An expected utility maximizer, as Becker, DeGroot and Marschak (1964) had shown, will set $\pi(X) = CE(X)$. Assuming that the subjects obey reduction, we can further obtain

$$A \sim R(A) = \left(\left(4, \frac{35\pi(X)}{360}; -1\frac{\pi(X)}{360} \right); \delta_{\pi(X)}, \frac{1}{1000}; \delta_{\pi(X)+0.01}, \frac{1}{1000}; \ldots; \delta_{9.99}, \frac{1}{1000} \right)$$

with *R(A)* corresponding to the tree in Figure 3.

Karni and Safra (1987) confirmed that if the independence principle is *not* obeyed, then it is not true that *always* setting $\pi(X) = CE(X)$

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

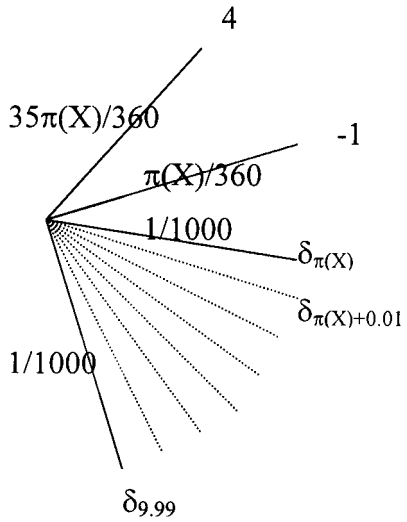

FIGURE 3. The lottery R(A)

maximizes the value of $R(A)$. A number of generalized theories which make do without independence (Karni and Safra call them 'Ω-theories', and I have given a partial list above) can in principle be applied to account for the data. Karni and Safra went further by putting forward an example of how PR were to be *expected* in the light of Quiggin's and Yaari's generalized model (EURDP), given a particular class of lotteries. The very pattern of choices observed by Lichtenstein and Slovic, Grether and Plott and others – as we shall see in Section 7 below – can be accounted for by applying EURDP to the BDM elicitation. If agents were EURDP maximizers, the data produced by means of the BDM mechanism would not be inconsistent with the transitivity of the underlying preferences, and the PR would be illusory. Karni and Safra (1987) showed also that a large class of BDM-like devices would be useless for eliciting non-linear preference relations.

By focusing on the reduction principle, Uzi Segal (1988) showed how violations of independence may not be the only causes of the PR 'illusion'. His argument, again, rests on the assumption that the agents perceive their task as a two-stage lottery,

$$\left(CE(X), \frac{\pi(X)}{10}; \lambda, 1 - \frac{\pi(X)}{10}\right),$$

where λ is a uniform distribution on the $[\pi(X), 9.99]$ interval. Segal's second step consists in conjecturing that agents do not satisfy reduction, and in conceiving an example constructed on particular pairs of bets: in some cases, again, a subject may price items in a way that would not reveal her true preferences.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

The general epistemic problem highlighted by the critiques of Holt, Karni and Safra, and Segal, is one of circularity. The 'instruments of observation' (elicitation) used in the experiments on individual choice rely heavily upon those theories of behaviour in whose investigation they are involved. Mechanisms such as the BDM procedure work by constructing further problems of choice under risk of the same kind as those being tested. There is clearly a problem of circular validation here: the phenomenon at stake is inconsistent with expected utility theory, but the instruments used to observe the phenomenon are constructed on the hypothesis that expected utility theory is correct. Is such a circle a vicious one? And if it is, how can it be escaped?

5. THEORY LADENNESS

According to the mild neo-positivism which shaped the standard view in philosophy of science, 'experimental laws' (the regularities produced in experimental situations – 'phenomena', in Bogen and Woodward's jargon) are independent of the truth of any particular scientific theory.[12] They constitute the neutral, solid bedrock upon which theories are constructed, and against which competing programmes are appraised. Such a view has been challenged by a number of authors, and it is nowadays fashionable to claim that observation reports in science are no less 'theory-laden' than high level explanations. The implications of such a thesis, however, are far from clear, because various different philosophical positions are often arbitrarily subsumed under the so-called 'theory ladenness of observation' label. Without aiming to have the last word on this complicated issue, then, let us try to make progress with it.

The expression 'theory-ladenness of observation' was introduced in the literature by Norwood Russell Hanson in his *Patterns of Discovery* (1958). Hanson suggested that the act of receiving a stimulus (visual, or otherwise) from the environment cannot be sharply distinguished from the act of attaching a meaning to it. There is no perception distinct from interpretation. The radical implications of this position were exploited a few years later by Thomas Kuhn in a famous chapter of his *Structure of Scientific Revolutions* (1962/1970, Chapter 10). Kuhn argued that scientists subscribing to different points of view interpret empirical data in the light of different theories, and therefore 'see different things' and 'seem to work in different worlds'. Thus, for example, 'the scientist who looks at a swinging stone can have no experience that is in principle more elementary than seeing a pendulum. The alternative is not some hypothetical "fixed" vision, but vision through an alternative paradigm,

[12] See Nagel (1961, p. 87).

one which makes the swinging stone something else' (1962/1970, p. 128).[13]

Similarly, in our case, PR constitute prima facie-evidence against the existence of well-ordered preference structures; but one may not trust what is 'seen' through the elicitation instruments because their functioning presupposes expected utility theory – a theory, moreover, for which there is established and extensive evidence of empirical violations. Taking theory ladenness seriously, one is invited to think that belief in the phenomenon rests on a sort of 'act of faith' regarding the theoretical framework used to observe the phenomenon. If this were the case, the proponents of different viewpoints would be condemned to argue in circles without ever reaching an agreement.

But is this the moral we should draw? As a matter of fact, the theory dependence of low-level scientific claims had already been noticed by pre-positivist philosophers of science. Classic statements can be found in Poincaré's and, most famously, in Pierre Duhem's writings. These authors stressed that experimental reports are by no means 'purely and simply an exposition of certain phenomena ['data', in Bogen and Woodward's jargon]; they are abstract propositions to which you can attach no meaning if you do not know the physical theories admitted by the author' (Duhem, 1906, p. 147). Surely the very language used in a scientific paper *does* involve theoretical presuppositions: how could we speak of 'currents' and 'potentials' without some theory of electricity, of 'risk-averse behaviour' without some theory of choice under uncertainty? Duhem, however, was not ready to go further than that. He was not willing to question the purity of sense-data, in particular: the rules for evaluating a scientist's report of what he has seen – 'in the strict sense of seeing with his own eyes' – are the same which apply to the testimony of a layman. Assuming he does not 'confuse the play of his imagination with perceptions' and masters his language, whenever the latter 'says that he has observed a fact, the fact is certain' (Duhem, 1906, pp. 158–9).

Henri Poincaré (1905) used to speak of 'crude facts' in a sense similar to Duhem's 'practical facts', as the neutral realm of everyday experience common to all sane human beings. According to Duhem and Poincaré – and contrary to what Hanson, Kuhn and their followers claim – there exists a level at which agreement on observational reports can in principle be reached. This seems to be true of the great majority of actual scientific controversies: data such as the position of a pointer on a scale, or the track on a bubble chamber picture are rarely the issues at stake. Scientists usually debate at a higher level of analysis than that, and it is not clear that the radical theory ladenness thesis can help us understand

[13] See also Kuhn's 'mature' position on the subject in his (1974), as well as Paul Feyerabend's (1975/1993, pp. 57–60) further elaboration.

how scientific controversies are resolved and why they are not. Experimental psychologists like Slovic and Lichtenstein see subjects choosing and pricing just like economists do, although they may not agree on what these subjects prefer (or even whether there is such a thing as preference, in economists' technical sense).

In the next section, I shall try to show that scientific disagreement normally has a two-fold origin: it is either reducible to a problem of (un-)reliability of the observation statements, or to a problem of fallibility of the assumptions and theories needed to infer from data to phenomena. In both cases, I shall argue, the problem can be framed in Duhemian form, and the disagreement can in principle be eliminated by means of further testing.

6. FALLIBILITY AND RELIABILITY

The first problem with observational reports has to do with the use of theoretical terms. Karl Popper (another anti-positivist sophisticated enough to notice that scientific claims 'are always interpretations of the facts in the light of theories', 1934/1959, Appendix *x) thought that even the simplest reports cannot be theory-free due to the ubiquity of what he called 'dispositional terms'. Universal names, 'words like 'glass' or 'water' are used to characterise the *law-like behaviour* of certain things' (Popper, 1934/1959, Appendix *x). But of course, one does not have to know that the term 'water' corresponds to 'a compound of two atoms of hydrogen and one of oxygen', when reporting that 'the value read on the scale of a thermometer immersed in ten litres of boiling water was 97.3 degrees Celsius'. It is worth, therefore, to draw a distinction between theoretical assumptions, in the sense of formal, explanatory high theory, and other non-rigorous presuppositions involved in the interpretations of facts. Whenever relevant, I shall write 'theory' to denote the former and 'background presuppositions' for the latter.[14] As Hacking (1983) has convincingly argued, knowledge of theory is seldom required in order to become a good 'observant', that is, a person particularly skillful at distinguishing certain patterns in a messy complex of sense-data.

What is the real problem raised by the presence of dispositional terms in observational language, then? Popper seems to be concerned with the *fallibility* of sentences including universal names: every such

[14] High theory is rarely – if ever – involved in reporting 'crude facts'. There are famous examples of data analysis done by laymen: for selecting the 290,000 bubble-chamber photographs taken at CERN during the experiments on weak neutral currents in 1973, non-experts were employed after a brief training. None of them knew the physics of small particles. The story of this experiment is told in Galison (1987). Cf. Hacking (1983, Ch. 12) for a taxonomy of different layers of theory typically involved in scientific activity.

statement, he says, 'has the character of a hypothesis' (1934/1959, §25). Was the animal we have just seen *really* a horse? We may find out tomorrow, for example, that our theory of the classification of animals is partly mistaken (as in the famous case of whales – formerly classified as fishes and then as mammals). This is nothing but the so-called Duhem problem of testing.[15]

It is well known that, according to Duhem, a scientific prediction can be made only by putting to work a 'whole theoretical scaffolding' (1906, p. 185). Whereas Duhem stressed the role of assumptions about the functioning of the instruments, it is customary today to interpret his thesis broadly, and include among the premisses used to deduce a prediction also presuppositions about the non-interferences of disturbing factors (or fulfilment of the *ceteris paribus* clause), the correct specification of the initial conditions, and so on.[16] When we seem to have produced a phenomenon contradicting our predictions, so the argument goes, we cannot *by deductive logic alone*, argue for the falsification of any one in particular of the assumptions involved (although we know that at least one must be false).[17] Duhem's problem can in the best cases be (partially) solved by checking each component of the inference at stake in other experimental circumstances, until every source of disagreement has been eliminated. Laboratory experimentation is particularly efficient in this regard, as we shall see in Section 8. For the time being, let us remember that *the problem of artefacts is mainly a problem of fallibility*.

A second problem concerns the *reliability* of our observations. Scientists, as a matter of fact, rarely disagree on observation reports; when that happens, however, their disagreement is rarely, if ever, a question of perceptual incommensurability. Most often, scientists doubt the correctness of others' reports, and look for mistakes in their data. For instance, the problem of data in the form of 'scintillations' (tiny flashes) counts was at the core of the famous Vienna–Cambridge controversy on protons and alpha-particles in the twenties.[18] The parties disagreed upon

[15] For some economic examples of Duhem's problem, the reader can see Cross (1982), Mongin (1988), and Sawyer, Beed and Sankey (1997).

[16] Cf. e.g. Lakatos (1970) and Putnam (1974). I intentionally refrain from interpreting Duhem's thesis as broadly as Quine (1953), according to which *the whole* system of one's belief (including the rules of deductive logic) is always at stake.

[17] In terms of propositional logic, the problem can be rendered as follows:
$(T \,\&\, IC \,\&\, A_1 \,\&\, \ldots \,\&\, A_n) \Rightarrow O$
$\sim O$

$\sim T \vee \sim IC \vee \sim A_1 \vee \ldots \vee \sim A_n$
where T is the theory under test, IC are the initial conditions, and $A_1 \ldots A_n$ are a stock of background assumptions of various sorts.

[18] See Stuwer (1985).

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

whether 'normal' observers could count the correct number of scintillations occurring during an experiment. But, as Bogen and Woodward (1988, p. 311) point out, this was a matter of reliability, not of theory-ladenness of data. There was here a clear matter of fact at stake – that is, the objective number of scintillations – and a solution to the controversy could have been reached, if one could just devise an independent and more reliable way to ascertain the fact of the matter.

Once again, problems of reliability are really, at the core, specific instances of Duhem's problem. The auxiliary assumption in question, in this case, is the proposition that 'all measurements have been performed correctly'. Putting it in terms of 'autopsychological' reports, for instance, the prediction 'When the water will be boiling, I shall read on the thermometer's scale a value of 100 degrees Celsius' can be deduced from a number of assumptions including the laws of thermodynamics, the assumption that the thermometer is not damaged, that the measurement has really been made at sea-level, that the water is pure, assumptions about the absence of major disturbances, and the proviso that I will not hallucinate when reading the thermometer. If I do not observe the predicted values, I shall not know automatically which of the assumptions to blame. But still, there will be some way to try and figure it out: by means of further testing.

Since they can be reduced to Duhem form, problems of fallibility and reliability of data are more tractable than problems of radical theory ladenness (if there exist in science any genuine examples of the latter). In principle, it should always be possible to 'extract' the relevant assumptions from the 'theory laden' data and add them as premises to make a larger Duhem problem.[19] When the issue is the truth of these assumptions, experiments can be devised to settle at least some disputes. The Cambridge–Vienna controversy was settled, among other things, by showing that human visual perception is unable to distinguish with the required precision different kinds of scintillations observed through the microscopes available at the time. It would have been possible *in principle* to write down the right report, it was just very difficult. Hanson's and Kuhn's point applies to those cases in which there is no 'right' way to see something, when there simply is no matter of fact to be ascertained. When the issue is reliability or fallibility, there is a way out of the corner: by means of successive tests we rule out some possibilities from the number of interpretations opened by the Duhem problem. In the next section I shall illustrate how this can be done in practice by illustrating some experiments on PR.

[19] Cf. Worrall (1991) for a similar argument directed against Feyerabend's version of the theory ladenness thesis.

7. INDEPENDENT TESTS

Neither Popper nor any other philosopher aware of Duhem's problem ever argued that a decision should be taken about which element (among the premises involved in a predictive argument) to revise without a reason being given. Their point was that such a reason cannot be an ultimate, logically compelling one. Relatedly, Otto Neurath (1934) noticed that in practice one does not have to deal with an infinite number of alternative interpretations of a given scientific observation. Testing is an obvious way to try to further reduce the number of plausible alternatives. The same idea can be found in Popper's writings: suppose that, as in the BDM case, someone challenges the standard inference from an observation to the falsification of a theory by putting forward a rival account. In order to test the legitimacy of such an alternative, one has to devise an independent test for it. A new experiment is needed because the old data do not provide a severe enough test of the new conjecture.

The notions of severity and independence of a test are strictly related to each other in the writings of Popper (see especially 1957 and 1963) and of his followers. Sometimes they are not distinguished carefully enough, but I shall not provide here an account of the various formulations of these two concepts throughout history.[20] Two ideas are relevant for us: (1) that when scientists have several alternative explanations of a certain event, they try as much as possible to reduce them in number; (2) that even when more than one explanation is left, it is rarely the case that the available evidence supports them all to the same degree – some theories will be *more* severely tested by the evidence than others.

In order to capture the second intuition, Elie Zahar (1976) and John Worrall (1978, 1985) have proposed a rather sophisticated criterion, according to which the facts which truly corroborate a theory have to be 'novel', but only in a relative sense: they may be already *known* at the time of the production of the theory, but *new for that particular theory*, that is, truly confirming phenomena must not have played a role in the construction of the theory at stake. Worrall and Zahar include among the corroborating evidence the data known at the time of the experiment, but which were not used to build the prediction at stake. Evidence cannot be used twice, in other words, once to construct and once to confirm.

We already said that Karni's and Safra's (1987) paper applies a specific generalized expected utility model – EURDP – to PR data. The way in which this is done suggests that the application is mostly illustrative. Certain passages indicate, however, that Karni and Safra

[20] This has already been done by Worrall (1978), and more recently Mayo (1996, Ch. 8).

believed that subjects do indeed violate independence and that this fact is the basis for an explanation of their behaviour.[21] But on what grounds?

Perhaps Karni and Safra were influenced by the general consensus achieved by Quiggin's and Yaari's EURDP among decision theorists. Generalized theories of decision making, however, do not imply that subjects violate independence – only that they might. Such theories display in their general form several free parameters which have to be fixed in order to derive precise implications about subjects' behaviour. For some special values of these parameters, the consequences of EURDP are identical to those of expected utility theory. In other words, EURDP may well be true and yet the subjects not violate independence when choosing among the lotteries typically used in PR experiments. Some *specific models* must be employed in order to account for PR – as Karni and Safra did in order to illustrate their main result. It is clear that such models needed independent confirmations before they could be taken seriously.

The models are obtained by '*ad hoc*' specification: the Karni and Safra reinterpretation of the BDM procedure holds, in fact, only for some pairs of lotteries and some values of the free parameters of the basic EURDP theory. According to the latter, the value V of a lottery $(x_1, p_1; \ldots; x_n, p_n)$ is given by

$$V(x_1,p_1;\ldots;x_n,p_n) = \sum_{i=1}^{n} u(x_i)\left[f\left(\sum_{j=i}^{n} p_j\right) - f\left(\sum_{j=i+1}^{n} p_j\right)\right].$$

The u is the traditional monotonic increasing real valued function defined on some interval in the real line (that is, on a range of monetary prizes). Compared to expected utility theory, EURDP has one more free parameter, namely the 'probability transformation function' f – a monotonic, increasing and continuous transformation from the unit interval onto itself. Karni and Safra (1987) show that *if* the following specifications are chosen for f and u,

$$f(p) = \begin{cases} 1.1564p, & 0 \leq p \leq 0.1833 \\ 0.9p + 0.047, & 0.1833 \leq p \leq 0.7 \\ 0.5p + 0.327 & 0.7 \leq p \leq 0.98 \\ p, & 0.98 \leq p \leq 1, \end{cases}$$

[21] For example: 'What Grether and Plott tried and – as our discussion indicates, *failed to do* – is to observe, by means of [the BDM method], the certainty equivalents of given lotteries' (Karni and Safra, 1987, p. 676, *own emphasis*). In a footnote Karni and Safra compare their contribution to Holt's: the latter pointed independently to violations of intransitivity, but 'however, did not present an alternative theory *explicating* the "PR" phenomenon' (*ibid*. p. 676, n4, *own emphasis*).

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

$$u(x) = \begin{cases} 30x + 30, & x \leq -1 \\ 10x + 10, & -1 \leq x \leq 12, \\ 6.75x + 49, & 12 \leq x, \end{cases}$$

then for lotteries such as the ones used by Grether and Plott (1979) – that is (−1, 1/36; 4, 35/36) and (−1.5, 25/36; 16, 11/36) – the 'announced price reversals' can be accounted for. (Notice: the 'announced price reversals' are the *data* to be explained, as opposed to the allegedly artefactual 'PR' *phenomenon*.) Still, these specifications are able to account only for the above lotteries: the Karni and Safra hypothesis cannot even rationalize all the data known at the time, unless one uses different parameter specifications for each experiment.[22] But even if this problem could be overcome, there would remain a general methodological concern. The illustrative *model*, with its particular parameters and initial conditions, rather than Quiggin's *theory* is doing most of the work. EURDP cannot by itself even account for the particular asymmetries of observed reversals: only the model with its specific parameters can. The illustrative model above (theory plus specification of the free parameters plus initial conditions) was in fact devised explicitly in order to account for the evidence to be explained. The latter, then, cannot provide much support to the violation of the independence hypothesis.

Following Zahar (1997), let us represent a theory with two free parameters (a_1 and a_2) as $T(a_1, a_2)$. A specific model of the theory can be devised by determining the free parameters on the basis of some empirical evidence e. Such evidence is used together with the theory T in order to deduce the values of the parameters: $[T(a_1, a_2)\ \&\ e] \Rightarrow T(a_1^*, a_2^*)$. The evidence e is in some cases, such as the present one, the very evidence the theory was intended to explain. It is not surprising, then, that e is accounted for by T.[23] Some new evidence e' not used to derive the model $T(a_1^*, a_2^*)$, but *implied* by $T(a_1^*, a_2^*)$, could confirm T more than e did, and even more so if no alternative theory is able to account for e'

[22] With hindsight, the popularity among economists of this explanation of PR appears puzzling (see also Hausman, 1992; Hausman and Mongin, 1998): why was Karni's and Safra's work so attractive? At least three explanations can be tentatively put forward: (1) the PR phenomenon looked so damaging to orthodox theory that economists were eager to believe more or less in *any* defensive argument whatsoever; (2) the strategy of weakening independence to account for counterexamples was then quite fashionable, therefore the Karni and Safra argument promoted theoretical unification; (3) the analysis of the BDM and RLS mechanisms had independent theoretical interest.

[23] Notice that e *weakly* confirms T, because it is logically conceivable that for some other theory $T_i \neq T$ there exists no set $\{a_1^*, \ldots, a_n^*\}$ such that $T_i(a_1^*, \ldots, a_n^*) \Rightarrow e$. In the limit case in which this were true for all T_i, T would be practically testable 'in isolation' and the Duhem problem would be drastically reduced. But, as in the case in hand, there usually exist a number of alternative theories able to account for e.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

(which will therefore count as a quasi-crucial experiment with respect to *T* and its rivals).

But why do certain data provide a better test of a hypothesis than others? According to Deborah Mayo (1996, Chapter 6), this does not have so much to do with novelty as with expectation.[24] Some data test a hypothesis severely only if it is very unlikely that those data can be produced while the hypothesis is false. The evidence *e* (the observed choice-price reversals) does not support Karni's and Safra's model because the classical PR experiment produces results consistent with that model regardless of the real truth-value of the latter. If we repeated the PR experiment even a thousand times, it would not be surprising to find that the resulting evidence is consistent with the Karni–Safra model – because the Karni–Safra model was constructed so as to accomodate the result of a PR experiment. The role of experiments is not to produce no-matter-what data. A good experiment must produce good data to answer specific questions, better than 'casual' data would. The original PR experiment was devised to answer Lichtenstein's and Slovic's question about the context-dependence of pricing and choice behaviour.[25] Other PR experiments (e.g., Grether's and Plott's) were performed later in order to answer other questions about artefacts. All these tests were designed so that it would have been really unlikely that a certain result (*e*) was observed if the tested hypothesis were false.[26] The data produced in those experiments cannot test Karni's and Safra's hypothesis severely. Some special experiment must be devised that is able to answer this new question in a convincing manner.

This reasoning can account for the behaviour of economic experimentalists, who soon began to design experiments to severely test the violation of the independence hypothesis. Subsequent work by Safra, Segal, and Spivak (1990b) was devoted to deriving further testable implications from the Karni–Safra interpretation of PR. According to Safra's, Segal's, and Spivak's *Proposition 2* (derived from Karni's and Safra's model), although the optimal selling price (π) of a lottery and its certainty equivalent (*CE*) may end up to be non-identical in a BDM

[24] Mayo (1996, Ch. 8) argues that her notion of severe test is able to account for all the positive intuitions of neo-Popperian theories of confirmation as well as to avoid some of their defects. The thrust of her argument against Worrall and Zahar is that there are cases in which data are used to construct a scientific hypothesis *and* at the same time provide a severe test of that hypothesis. The interested reader can look in particular at Sections 8.3–8.4 of her book.

[25] See Section 2 above.

[26] For instance, we can formulate one of the (low-level) hypotheses tested by Grether and Plott (1979) as '*e* is not due to the absence of relevant monetary rewards'. By performing experiments with relevant monetary rewards, Grether and Plott constructed a severe test of their hypothesis, by making it unlikely that *e* was produced in their experiment if the hypothesis was false.

elicitation, they should nevertheless lie on the same side of the lottery's expected value (EV). In other words, the two following testable predictions (for risk-loving and risk-averse subjects respectively) can be derived from Karni's and Safra's analysis:[27]

(i) $CE(X) > EV(X) \Rightarrow \pi(X) \geq EV(X)$
(ii) $CE(X) < EV(X) \Rightarrow \pi(X) \leq EV(X)$

According to Segal's (1988) interpretation of the BDM device, on the other hand, (i) and (ii) do not necessarily hold. An experiment testing such predictions would certainly count as an independent test of Karni's and Safra's hypothesis. Keller, Segal and Wang (1993) ran such an experiment, and found 'Proposition 2' to be inconsistent with around 30% of the data. Such a percentage cannot be explained as random error, because a definite asymmetric tendency is discernible in the data: the $\pi(X) > EV(X) > CE(X)$ pattern is displayed for 22% of the subjects, whereas the $CE(X) > EV(X) > \pi(X)$ pattern is shown for a 9% only (Camerer, 1995, p. 659). This seems to rule out the explanation of PR in terms of violations of independence definitively, leaving open the issue whether price reversals are effects of reduction violations (as conjectured by Segal) or symptoms of intransitive preferences.

Safra, Segal and Spivak (1990a) also proved that the kind of explanation that Karni and Safra gave of the PR phenomenon presupposes similar conditions to those known to be at the origins of other anomalies like Allais's paradox and the so-called common ratio effect. All these phenomena (including PR) ought therefore to be explicable in Mark Machina's 'Generalized Expected Utility Analysis' (1982), by assuming the so-called 'fanning out' hypothesis.[28] MacDonald, Huth, and Taube (1991) devised an experiment to test the 'violation-of-independence' explanation of PR by checking whether there is a correlation between fanning out and reversals. They ran Allais-type experiments and then PR experiments, and found that subjects who did *not* exhibit fanning out did *not* in fact incur fewer reversals than the others.

Another example of the same kind is provided by Starmer's and Sugden's (1991) attempt to test Holt's hypothesis.[29] Starmer and Sugden

[27] For the technical details of such a derivation, cf. Safra, Segal and Spivak (1990b, pp. 187–8). Notice that (i) and (ii) can be derived also from expected utility theory, since according to the latter $CE(X) = \pi(X)$.

[28] I.e., that a local utility function displays greater risk-aversion than another local utility function, if the first is defined at a lottery which stochastically dominates the lottery at which the second is defined.

[29] To be precise, Holt (1986) does not rule out that PR data be an effect of reduction violation (he actually cites explicitly this event, p. 514); yet, when it comes to illustrating a possible

designed an experiment in which the reduction hypothesis, upon which Holt's main example apparently depends, is incompatible with a very frequent violation of independence first discovered by Allais (1953/1979), the 'common consequence' effect. The Allais-type experiment involved a double choice, first between a lottery R' = (£10, 0.2; £7, 0.75; 0, 0.5) and a lottery S' = (£7, 1); then between R'' = (0, 0.8; £10, 0.2) and S'' = (£7, 0.25; 0, 0.75). The common consequence effect is a tendency to choose $S' > R'$ and $R'' > S''$. By reduction, it is easy to show that the following equivalence holds between compound lotteries: (R', 0.5; S'', 0.5) = (S', 0.5; R'', 0.5) = (£10, 0.1; £7, 0.5; 0, 0.4). If there is reduction, then, one should expect a random pattern of choice between (R', 0.5; S'', 0.5) and (S', 0.5; R'', 0.5) whereas common consequence implies (S', 0.5; R'', 0.5) > (R', 0.5; S'', 0.5). If there is reduction, in other words, there cannot be common consequence effects, and vice-versa. Starmer and Sugden performed the above Allais-type experiment with and without the RLS mechanism, and observed the same ratio of common consequence violations in all cases. This provided strong evidence that subjects did not obey reduction, and the 'Holt hypothesis' was discredited.

These experiments have reinforced scientists' belief in the reality of the PR phenomenon by elimination of alternative explanations. Attempts to explain away the phenomenon have been rejected by testing the various alternative hypotheses independently. In all these cases, experimental economics seems to follow a classic logic of theory-testing: a theory is proposed; predictions are derived from it (plus initial conditions and auxiliary hypotheses); finally, such predictions are checked against empirical data, and either the assumptions are rejected, or they are confirmed. There are, however, other means to increase our belief in the reality of a phenomenon or to reduce the plausibility of artificiality claims. These have to do not so much with the classic scheme of hypothesis testing, but rather take the form of 'observations without (a unique) theory'. I shall turn to these in the next section.

8. PHENOMENA WITHOUT (A UNIQUE) THEORY

Traditional philosophy of science tended to portrait all experiments as tests of some theoretical hypothesis. In contrast, recent studies of experiments have tried to define boundaries to the importance of high-level theory in experimental science. As a matter of fact, experiments like those on PR do not have primarily to do with any explanatory theory. Their motivation may rather be summarized as follows: how can we

mechanism to account for PR data, Holt devises an example based on reduction and violation of independence. I shall follow the literature and refer to this illustration as to the 'Holt hypothesis'.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

know that the observations of preference relations made via elicitation mechanisms were reliable?

To appreciate this distinction, consider the experiments discussed at the end of the previous section. Starmer and Sugden tried to test the general validity of Holt's reduction hypothesis, and thus their experiment was primarily a theory-testing one. But the fact that a statistically significant portion of subjects violates reduction does not prove that the RLS is a reliable procedure to elicit preferences. There may still be a significant portion of subjects behaving the Holt way – that is, the reduction hypothesis may be false *in general*, and at the same time the RLS procedure be unrealiable *in general*. To account for this, one has to devise an experiment able to check the capacity of the RLS mechanism to elicit true preferences. Beattie and Loomes (1997) and Cubitt, Starmer and Sugden (1998) have tried to calibrate the data obtained by means of RLS against single-choice experiments in which each task is played (and paid) 'for real'. Experiments of this sort are costly and therefore rare, but the available evidence seems to corroborate the reliability of the RLS mechanism at least for simple choices.

The BDM mechanism should in principle be submitted to the same kind of investigation. Edi Karni and Zvi Safra posed the following two questions: (a) 'How rich is the class of preferences that permits the elicitation of certainty equivalents of given lotteries using [the BDM] method?'; and (b) 'Are there experiments that enable the elicitation of the certainty equivalents of every lottery for every reasonable preference relation?'. The first question is the one that motivated their enterprise; the second one is rather more ambitious. The answers provided by Karni and Safra are respectively that '(a) the elicitation of certainty equivalents of all lotteries, using the experimental methods of Becker, DeGroot, and Marschak, is possible if and only if the preference relation is representable by an expected utility functional; (b) every experiment in a larger class of experiments [which Karni and Safra call 'Q-experiments'] would fail to elicit the certainty equivalent of some lotteries for some reasonable preference relations' (Karni and Safra, 1987, p. 676). In other words, if subjects' decisions violate independence in the cases at hands, then the BDM procedure and similar mechanisms are not adequate instruments to determine certainty equivalents *in all cases*. From this, Karni and Safra conclude that 'Grether and Plott and others [...], as our discussion indicates failed to [...] observe by means of an experimental method developed by Becker, DeGroot and Marschak (1964), the certainty equivalents of given lotteries' (*ibid*.). However, it does not seem to follow logically that PR cannot be detected by means of 'Q-experiments'. The BDM mechanism and similar methods may not be universally precise, but still precise enough to establish that PR take place. One natural way to see whether this is true is to try to observe reversals with and without

the BDM procedure, and check whether it makes any difference. Let us see more precisely how this can be done.

In the case of PR, the phenomenon had been observed right from the beginning with and without elicitation mechanisms. Of Lichtenstein's and Slovic's early experiments (1971), only two involved the BDM procedure, but reversals were produced in all tests. This fact should have already been a puzzle to the Holt, Segal, Karni and Safra explanations. Years later, several experimenters began looking for a way to reproduce PR with incentive mechanisms but avoiding possible problems with the BDM procedure. The main obstacle was Karni's and Safra's argument that it is very difficult to construct an elicitation mechanism performing the same task as the BDM procedure without stumbling into the same problems with the independence axiom.

After a first dubious attempt by Cox and Epstein (1989), Amos Tversky, Paul Slovic and Daniel Kahneman (1990) managed to create an incentive procedure able to elicit orderings without the BDM mechanism. The subjects were initially required to price all the lotteries in random sequence; only later, the bets were paired and the subjects asked to choose between them. One pair was picked up by RLS, and then one bet among the highest priced and the chosen one was randomly selected and played. Not knowing the composition of each pair at the moment of pricing, the subjects were incited to state their true selling price in order to ensure that their preferred one was always picked up and played. Under this design, the only pattern of reversals that can be rationalized by a generalized model without the independence axiom is a random one – when subjects prefer a fifty-fifty chance of playing either lottery in a pair to playing one of them for sure. The pattern of PR observed by Tversky, Slovic and Kahnemann was too close to the standard one to be interpreted as random behaviour.

Remember the target of the arguments of Holt, Karni, Safra and Segal: they intended to show that it was not intransitive preferences that had been observed in experiments with the BDM procedure. In order to reject Karni's and Safra's and Segal's interpretation, therefore, one must not necessarily show that their theories are erroneous. It should be sufficient to show that it was really a feature of preferences that was observed in the experiments in question. Hacking, in a chapter entirely devoted to discussing the reliability of vision through microscopes (1983, Ch. 11), argues that powerful support to belief in the reality of what we see through electron microscopes is provided by the fact that the same structures are observed through *light* microscopes. The intuition behind this inference is captured by a so-called 'no-miracles' argument: it would really be a 'preposterous coincidence' if two instruments based on two entirely independent physical mechanisms ended up delivering two *identical*, and *artefactual*, sets of data. Our belief in the reality of a

phenomenon, notice, can be totally independent of the explanation we give of such a phenomenon. We may not know the causes of the phenomenon, nor have an established theory of the instrument, and yet believe in the phenomenon and in what we 'see' through the experimental apparatus.[30]

Finally, some experimental strategies involve a more substantial use of theory in supporting the reality of a phenomenon – but do so in order to justify the adoption of the RLS–BDM machinery rather than to explain the phenomenon. MacDonald's, Huth's and Taube's (1992) results suggested that a so-called 'isolation effect' may have been present in their experiments. Subjects, in other words, seemed to choose as though they evaluated each lottery in isolation, without multiplying its chances for the probability of its being selected by the RLS procedure. Such a procedure would be consistent with theories of decision, like Tversky's and Kahneman's 'Prospect Theory' (1979), which assume a desire on subjects' part to minimize the computational costs of decision making. MacDonald, Huth and Taube devised an experiment where subjects were offered the chance to revise their choice after a RLS was performed at the end of the experiment. A subject not obeying independence should change his choice, but very few did change despite the fact that strong independence violations were observed in previous experiments. Thus, some isolation effect may have been there. Isolation effects have the interesting property of counterbalancing the effect of independence violations: the BDM procedure, remember, fails to elicit true certainty equivalents only if independence is violated *and* reduction is obeyed. If subjects violate *both* independence and reduction by isolation, then the BDM machine may work well (as suggested by Camerer, 1989).

9. THE REALITY OF REVERSALS

Colin Camerer, in a recent survey of the PR experiments, concludes that the PR phenomenon can hardly be considered an artefact of the instruments of observation. That is today the standard view. Of the arguments he cites in its support, two are from refutation of alternative explanations: (a) from failed predictions derived from generalized expected utility models; and (b) from the evidence of 'isolation effects'.

[30] Jean Perrin's determination of Avogadro's number is another famous example of 'no-miracles' argument. In order to be sure that he had found the true value, Perrin checked his results by measuring it in alternative ways. In his 1913 book, *Les atomes*, Perrin reports thirteen different independent methods to ascertain Avogadro's number. The 'miraculous' convergence of all measures was taken to be an extremely strong proof that the result obtained was not an artefact of the procedures he had used. Even a conventionalist like Poincaré was struck by such a result; see Nye (1972) for the full story. For other examples, cf. Franklin (1986, pp. 131–5); for a philosophical discussion, see Salmon (1984, p. 216) and Mayo (1996).

The third argument (c) starts from the recognition that the same phenomenon appears to be identifiable via different mechanisms relying on different principles (Camerer, 1995, p. 659). Experimental economists may not know exactly how the BDM mechanism works (although they surely understand it better now than ten years ago), but they are confident today that it may still be used to detect PR. Experimentalists rely on several resources to establish the reality of a phenomenon. Allan Franklin (1990) lists nine possible strategies adopted by scientists in order to provide 'reasonable belief in the validity of an experimental result' (p. 104):

1. Experimental checks and calibration, in which the apparatus reproduces known phenomena.
2. Reproducing artifacts that are known in advance to be present.
3. Intervention, in which the experimenter manipulates the object under observation.
4. Independent confirmation using different experiments.
5. Elimination of plausible sources of error and alternative explanations of the result.
6. Using the results themselves to argue for their validity.
7. Using an independently well-corroborated theory of the phenomena to explain the results.
8. Using an apparatus based on a well-corroborated theory.
9. Using statistical arguments.

The strategies we have been concerned with in this reconstruction of the debate on PR belong to the italicized categories one, four and five above. Thanks to such strategies, the debate on the existence of PR was in effect brought to an end and disagreement eventually eliminated. This, it must be stressed, is quite a remarkable event in economics, where controversies tend to last for decades without parties ever converging on a common position. Such a result was made possible by the use of controlled experimentation. Other strategies in Franklin's list were clearly not available to economic experimentalists in the case in hand: the phenomenon under study, for instance, was inconsistent with accepted economic theory, which therefore could not be used to increase economists' confidence in the reality of PR. Most efforts are nowadays devoted to discovering whether the PR phenomenon is robust at a level higher than change of elicitation procedure, that is, whether it can be observed in the context of real market behaviour. This activity suggests some closing remarks on the notion of 'artefact'.

10. TWO KINDS OF ARTEFACTS

The causes of PR have not been established yet, and recent investigations invite the thought that reversals may result from the interaction of a

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

number of heterogeneous factors – but PR are very likely to be real.[31] The *Oxford English Dictionary* defines an 'artefact' as 'something observed in a scientific investigation, experiment, etc., that is not naturally present but originates in the preparative or investigative procedure or extraneously'. It is a different concept of artefact from the one discussed so far. When speaking of artefacts, as we have seen, often scientists want to suggest that they do not occur at all: for example, that whereas announced price reversals are real, PR are not.

PR are not artefacts in this latter narrow sense, and yet they might be artificial in the sense of 'not occurring naturally or spontaneously'. According to the *OED*, the first definition of 'artificial' is 'made by or resulting from art or artifice; constructed, contrived; not natural (though real)'.[32] Of course such a definition is unsatisfactory too: men are, after all, part of nature, and so are the objects they produce. The dichotomy 'spontaneous vs. forced' looks just as unsatisfactory as the 'natural vs. artificial' one: no phenomena occur spontaneously (independently of human intervention) in an economy, because all are (perhaps unintentional) effects of the intentional actions of a number of human beings. Therefore, what does 'artificial' mean in this case? Vernon Smith puts it in the following way:

> Once replicable results have been documented in laboratory experiments, one's scientific curiosity naturally asks if these results also apply to other environments, particularly those of the field. Since economic theory has been inspired by field environments, we would like to know, if we were lucky enough to have a theory fail to be falsified in the laboratory, whether our good luck will also extend to the field. Even if our theories have been falsified, or if we have no theory of certain well-documented behavioural results in the laboratory, we would like to know if such results are transferable to field environments. (1982, p. 267)

According to Smith, economic theory is concerned with a certain realm of social phenomena, namely market ones. Most such phenomena occur in circumstances that are rather different from those experimental situations where PR have been first produced and then studied. Such experimental environments are 'artificial' in the sense of being created especially for a certain scientific purpose. It is therefore important to check whether an experimental claim can be generalized to situations lying outside the (laboratory) domain where the phenomenon has first been observed.

[31] On the possible causes of PR, see Camerer's survey (1995). Cf. also Tversky, Slovic and Kahneman (1990) and Slovic (1995) for a statement of the point of view of psychologists on the matter.

[32] Cf. Hacking (1988, p. 285) for a discussion of the notion of 'artefact' which follows a similar line of reasoning.

TABLE 1. Different kinds of artefact

	Real	*Unreal*
Instantiated in the intended domain of economics	Genuine economic phenomenon	Artefact_1
Not instantiated in the intended domain of economics	Artefact_2	Artefact_1

In order to clarify the terminology, let us speak of artificial phenomena as either 'artefacts$_1$' or 'artefacts$_2$'. Artefacts$_1$ are misinterpreted data, mere 'illusions of phenomena'. Artefacts$_2$ are phenomena, but may be artificial nonetheless in a different sense.

For instance Holt, Segal, Karni and Safra concluded at the time that PR did not exist at all; they were 'illusions' produced by a mistaken theory of the instruments of observation (Table 1, right-hand boxes). The experiments reviewed in this paper have taken care of this interpretation (left-hand boxes); but they have not shown that PR are phenomena falling in the 'intended domain' of economics (upper left-hand box). The PR phenomenon may be 'non-genuine' just like an artificial heart is not a real heart (lower left-hand box). Does the phenomenon occur in 'real' markets? These constitute the intended domain of economic theories, and the relevance of Lichtenstein's and Slovic's discovery seems to depend very much on this issue. The present challenge consists therefore in showing that the artificial systems created in the laboratory are a good replica of (at least some) real economic systems.[33]

In *The Republic*, Plato distinguishes between the art of the craftsman (for example, the shoemaker or the carpenter) and that of the artist. Both produce artefacts, but of a different kind: whereas the former *makes* a craft (a shoe or a bed) which is an imperfect but concrete reproduction of the ideal (the 'bed in itself'), the latter *represents* the craftsman's product, thus creating 'mere phenomenal appearance' or 'illusion' (Book 10, 598b).[34] The artefacts we have been mainly concerned with in this paper are not unlike Plato's artworks. If the actual working of markets is the unknown ideal, these artefacts are, like artworks, 'at third remove from the throne of the truth' (Book 10, 597e).

[33] Experimenters have already started to tackle this issue. I review some experiments and discuss their methodological implications in my (1999) article.

[34] I am using Desmond Lee's translation, from the 1974 Penguin edition.

https://doi.org/10.1017/S0266267100000134 Published online by Cambridge University Press

REFERENCES

Allais, Maurice. 1953/1979. 'The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School'. In *Expected Utility Hypothesis and the Allais Paradox*. M. Allais and O. Hagen (eds.). Reidel

Beattie, Jane and Graham Loomes. 1997. 'The impact of incentives upon risky choice experiments'. *Journal of Risk and Uncertainty*, 14:149–62

Becker, Gary M., Morris H. DeGroot, and Jacob Marschak. 1964. 'Measuring utility by a single-response sequential method'. *Behavioral Science*, 9:226–32

Bogen, James and James Woodward. 1988. 'Saving the phenomena'. *Philosophical Review*, 97:303–52

Camerer, Colin. 1989. 'An experimental test of several generalised utility theories'. *Journal of Risk and Uncertainty*, 2:61–104

Camerer, Colin. 1995. 'Individual decision making'. In *The Handbook of Experimental Economics*. J. H. Kagel and A. E. Roth (eds.). Princeton University Press

Chew, S. H. and K. MacCrimmon. 1979. Alpha-nu choice theory: a generalization of expected utility theory. University of Columbia, Faculty of Commerce and Business Administration, Working Paper no. 686

Collins, Harry M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. Sage

Cox, James C. and Seth Epstein. 1989. 'Preference reversals without the independence axiom'. *American Economic Review*, 79:408–26

Cross, Rod. 1982. 'The Duhem-Quine thesis, Lakatos and the appraisal of theories in macroeconomics'. *Economic Journal*, 92:320–40

Cubitt, Robin P., Chris Starmer and Robert Sugden. 1998. 'On the validity of the random lottery incentive system'. *Experimental Economics*, 1:115–31

Duhem, Pierre. 1906. *La théorie physique. Son objet et sa structure*. Chevalier et Rivière. Engl. trans. *The Aim and Structure of Physical Theory*. Princeton University Press. 1954

Feyerabend, Paul K. 1975/1993. *Against Method*. Verso

Franklin, Allan. 1986. *The Neglect of Experiment*. Cambridge University Press

Franklin, Allan. 1990. *Experiment, Right or Wrong*. Cambridge University Press

Galison, Peter. 1987. *How Experiments End*. University of Chicago Press

Grether, David and Charles Plott. 1979. 'Economic theory of choice and the preference reversal phenomenon'. *American Economic Review*, 69:623–38

Guala, Francesco. 1999. 'The problem of external validity (or 'parallelism') in experimental economics'. *Social Science Information*, 38:555–73

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge University Press

Hacking, Ian. 1988. 'The participant irrealist at large in the laboratory'. *British Journal for the Philosophy of Science*, 39:277–94

Hanson, Norwood R. 1958. *Patterns of Discovery*. Cambridge University Press

Hausman, Daniel M. 1992. *The Inexact and Separate Science of Economics*. Cambridge University Press

Hausman, Daniel M. and Philippe Mongin. 1998. 'Economists' responses to anomalies: full cost pricing versus preference reversals'. In *New Economics and Its History*. J. Davis (ed.). Supplement to *History of Political Economy*, Vol. 29. Duke University Press

Holt, Charles A. 1986. 'Preference reversals and the independence axiom'. *American Economic Review*, 76:508–15

Kahneman, Daniel and Amos Tversky. 1979. 'Prospect theory: an analysis of decision under risk'. *Econometrica*, 47:263–91

Karni, Edi and Zvi Safra. 1987. ''"Preference reversal" and the observability of preferences by experimental methods'. *Econometrica*, 55:675–85

Keller, L. Robin, Uzi Segal, and T. Wang. 1993. 'The Becker–DeGroot–Marschak mechanism and generalized utility theories: theoretical predictions and empirical observations'. *Theory and Decision*, 34:83–97

Kuhn, Thomas S. 1962/1970. *The Structure of Scientific Revolutions*. University of Chicago Press

Kuhn, Thomas S. 1974. 'Second thoughts on paradigms'. In *The Essential Tension*. University of Chicago Press

Lakatos, Imre. 1970. 'Falsificationism and the methodology of scientific research programmes'. In *Criticism and Growth of Knowledge*. I. Lakatos and A. Musgrave (eds.). Cambridge University Press

Lichtenstein, Sarah and Paul Slovic. 1971. 'Reversals of preference between bids and choices in gambling decisions'. *Journal of Experimental Psychology*, 89:46–55

Lichtenstein, Sarah and Paul Slovic. 1973. 'Response-induced reversals of preference in gambling: an extended replication in Las Vegas'. *Journal of Experimental Psychology*, 101:16–20

Loomes, Graham, Chris Starmer and Robert Sugden. 1989. 'Preference reversal: information-processing effect or rational non-transitive choice?'. *Economic Journal*, 99:140–51

MacDonald, D. N., William L. Huth, and Peter M. Taube. 1992. 'Generalized expected utility analysis and preference reversals: some initial results in the loss domain'. *Journal of Economic Behavior and Organization*, 17:115–30

Machina, Mark J. 1982. '"Expected utility" analysis without the independence axiom'. *Econometrica*, 50:277–323

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. University of Chicago Press

Mongin, Philippe. 1988. 'Problèmes de Duhem en théorie de l'utilité espérée'. *Fundamenta Scientiae*, 9:299–327

Nagel, Ernst. 1961. *The Structure of Science*. Harcourt, Brace & Wold

Neurath, Otto. 1934. 'Radical physicalism and the "real world"'. In *Philosophical Papers 1913–1946*. R. S. Cohen and M. Neurath (eds. 1983). Reidel

Nye, Mary J. 1972. *Molecular Reality*. Macdonald

Poincaré, Henri. 1905. *La valeur de la science*. Flammarion. Engl. trans. *The Value of Science*. Dover Publications

Pommerehne, Werner W., Friedrich Schneider and Peter Zweifel. 1982. 'Economic theory of choice and the preference reversal phenomenon: a reexamination'. *American Economic Review*, 72:569–74

Popper, Karl R. 1934/1959. *Logic of Scientific Discovery*. Hutchison

Popper, Karl R. 1957. 'The Aim of science'. *Ratio*, 1:24–35. Reprinted in *Objective Knowledge*. Clarendon Press

Popper, Karl R. 1963. *Conjectures and Refutations*. Routledge

Putnam, Hilary. 1974. 'On the "corroboration" of theories'. In *The Philosophy of Karl Popper*. P. Schilpp (ed.) Open Court

Quiggin, John. 1982. 'A theory of anticipated utility'. *Journal of Economic Behavior and Organization*, 3:323–43

Quine, Willard O. 1953. 'Two dogmas of empiricism'. In *From A Logical Point of View*. Harvard University Press

Reilly, Robert J. 1982. 'Preference reversal: further evidence and some suggested modifications in experimental design'. *American Economic Review*, 72:576–84

Safra, Zvi, Uzi Segal and A. Spivak. 1990a. 'Preference reversals and non-expected utility'. *American Economic Review*, 80:922–30

Safra, Zvi, Uzi Segal, and A. Spivak. 1990b. 'The Becker–DeGroot–Marschak mechanism and non-expected utility: a testable approach'. *Journal of Risk and Uncertainty*, 3:177–90

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press

Sawyer, K. R., Clive Beed and Harold Sankey. 1997. 'Underdetermination in economics. The Duhem-Quine thesis'. *Economics and Philosophy*, 13:1–23

Segal, Uzi. 1988. 'Does the preference reversals phenomenon necessarily contradict the independence axiom?'. *American Economic Review*, 28:175–202

Slovic, Paul. 1995. 'The construction of preferences'. *American Psychologist*, 50:364–71

Slovic, Paul and Sarah Lichtenstein. 1968. 'Relative importance of probabilities and payoffs in risk-taking'. *Journal of Experimental Psychology*, Supplement, Part 2:1–18

Smith, Vernon L. 1982. 'Microeconomic systems as an experimental science'. *American Economic Review*, 72:923–55

Starmer, Chris and Robert Sugden. 1991. 'Does the random-lottery incentive system elicit true preferences? An experimental investigation'. *American Economic Review*, 81:971–8

Stuwer, R. H. 1985. 'Artificial disintegration and the Vienna–Cambridge constroversy'. In *Observation, Experiment and Hypothesis in Modern Physical Science*. P. Achistein and O. Hannaway (eds.). MIT Press

Tammi, Timo. 1999. 'Incentives and preference reversals: escape moves and community decisions in experimental economics'. *Journal of Economic Methodology*, 6:372–402

Thaler, Richard H. and Amos Tversky. 1990. 'Preference reversals'. *Journal of Economic Perspectives*, 4:201–11

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. 'The causes of preference reversals'. *American Economic Review*, 80:204–17

Worrall, John. 1978. 'The ways in which the methodology of scientific research programmes improves on Popper's methodology'. In *Progress and Rationality of Science*. G. Andersson and A. Radnitzky (eds.). Reidel

Worrall, John. 1985. 'Scientific discovery and theory-confirmation'. In *Change and Progress in Modern Science*. J. Pitt (ed.). Reidel

Worrall, John. 1991. 'Feyerabend and the facts'. In *Beyond Reason: Essays on the Philosophy of Paul K. Feyerabend*. G. Munévar (ed.). Kluwer

Yaari, Menachim E. 1987. 'The dual theory of choice under risk'. *Econometrica*, 55:95–115

Zahar, Elie. 1976. 'Why did Einstein's programme supersede Lorentz's'. In *Method and Appraisal in the Physical Sciences*. C. Howson (ed.). Cambridge University Press

Zahar, Elie. 1997. *Leçons d'épistemologie*. In *Cahiers du CREA*. A. Boyer (ed.). École Polytechnique