self-experimentation – most notably from psychoanalysis, which started with Freud's self-analysis.

We also have several statistical concerns. Roberts claims that single-subject experiments do not require different inferential statistical analysis than typical, multiple-subject experimental and observational research. We object to this claim and approach to the analysis. It is contradictory that Roberts favors visual display of data and exploratory data analysis, citing Tukey's (1977) seminal work, and at the same time reports numerous inferential statistical tests. On the other hand, the target article is silent about effect size indicators of treatments or interventions (Cohen 1988). Effect sizes are critical for exploratory data analysis and should be included. Also, the overuse of the loess procedure, locally weighted regression (Cleveland 1993), is problematic as the data are often noisy and may lead to dubious conclusions.

Further, there is a serious omission concerning the smoothing of time-series data, and appropriate time-series analyses are never performed. The loess procedure is not appropriate for time-series data which are necessarily serially correlated. Specifically, the loess procedure is a regression technique for serially uncorrelated data with independent realizations of a random variable.

Roberts' contribution represents 10 years of self-experimentation, and yet we remain uninformed about its reliability, confounds, generality, interindividual differences, and causality. Unfortunately, this approach is comprised solely of single-subject observations over time, and thus tells us nothing about the variability among persons that is the core of psychology. Nevertheless, the target article does incorporate one extremely useful reminder: Data is an encyclopedia for future research.

Introspection and intuition in the decision sciences

Daniel John Zizzo

Department of Economics, University of Oxford, Oxford, OX1 3UQ, United Kingdom. **daniel.zizzo@economics.ox.ac.uk**
http://www.econ.ox.ac.uk/Research/Ree/ZizzoWebpage.htm

Abstract: Self-experimentation is uncommon in the decision sciences, but mental experiments are common; for example, intuition and introspection are often used by theoretical economists as justifications for their models. While introspection can be useful for the generation of ideas, it can also be overused and become a comfortable illusion for the theorist and an obstacle for science.

This commentary complements the target article by analyzing the role that self-experimentation and mental experiments play in research on decision-making. Whereas Roberts appears to claim that participation in one's own experiments is common in human experimental psychology, my expectation is that, as a rule, papers in the decision sciences fail to have the experimenters as part of the sample. Of course, exceptions exist, especially in consumer research (e.g., Earl 2001), although even then they are controversial and usually thought of in terms of researcher introspection (Gould 1995; Wallendorf & Brucks 1993). They also tend to be less systematic than Roberts' own work (e.g., Earl [2001] is just a case study).

Roberts' target article makes sufficiently clear that *self-experimentation* means to run an experiment with oneself as a subject, going through one or more experimental conditions on the basis of a carefully drawn experimental design. In the decision sciences, the researcher would gather behavioral data, such as those one could get from other subjects, but presumably in a more flexible way (one of the advantages mentioned by Roberts); she may also collect nonbehavioral data, the reliability of which lies in introspection, and which may be richer with herself as a subject. While self-experimentation is uncommon, *mental experiments* are common and even pervasive: the experimenter considers what choices she (or some other economic agents) would make in a hypothetical situation. Mental experiments "work" insofar as introspection works. As recognized by Hutchison (1938), they have a role to play in the generation of scientific ideas (Earl 2001), which fits well with the target article. Introspection was used to defend the concept of cardinal utility in the first half of the twentieth century (Lewin 1996). Even today, economists all too often refer to "intuition" and "reasonableness" as the great virtues of their theoretical models of decision-making, thus calling for their readers, and students, to rely on their introspection (e.g., Varian 1999, p. 3).

Unfortunately, introspection has limits. First, people differ in cognitive abilities and social preferences, and population heterogeneity cannot be captured in an experiment with size $n = 1$. Second, introspection may be marred by various judgement biases, such as confirmation and hindsight bias (see Camerer 1995). For example, Dawes and Mulford (1997) showed that conjunction effects in memory recollection may explain some purported evidence for alien kidnappings (conjunction effects occur whenever an agent ranks the conjunction of two events as more likely than the less likely of the two events: Stolarz-Fantino et al. 2003). Third, there is a dissociation between what people know and learn explicitly (e.g., the laws of physics enabling people to ride bikes) and what they can actually do and learn implicitly (e.g., actually riding a bike; see Shanks & St. John 1994). For example, Zizzo (2003) found a dissociation between verbal and behavioral learning in a conjunction effect task: verbal responses were sensitive, but actual behavioral choices were entirely insensitive to the amount of verbal instructions being provided. Introspection may give access to the first source of knowledge and learning, but not to the second. Introspection may give no answers, or the wrong answers, where implicit knowledge and learning are important; for example, in the conjunction effect task example, it may lead a researcher to think that with the right feedback the effect may disappear, even though it would not in relation to behavioral choices.

Fourth, the potential dependence of behavior on *explicit* knowledge is *also* a problem. Assume that a decision scientist said "people never commit conjunction fallacies" and as evidence she brought the fact that she does not commit them. The problem is obvious: the evidence for *P* is weak if the agent can cause *P* because of her beliefs about *P.* Elster (1983) made an important distinction between primary and secondary states. Primary states are what an agent can directly choose (one can go to bed at 10 p.m.); secondary states may occur as a by-product of other states, including primary states, but are not directly chosen (an agent may go to bed at 10 p.m., and this may help her get to sleep soon afterwards, but she cannot *choose* to fall asleep at exactly 10 p.m.). For secondary states, while concerns about expectations and placebo effects do exist, the link they have with the experimental variable being investigated is indirect. Conversely, in the case of primary states, a researcher directly chooses the value of the experimental variable: for example, she chooses what action to take in a social dilemma, or which lottery to play out if she is asked to pick one of two. This creates difficulties. First, whether consciously or not, having a theory about *P* may modify a researcher's behavior to be more congruent with *P.* More important, to the extent that *P* is ruled by one's own explicit knowledge, *P will* be acted out if *P* has normative value, for example, the researcher believes that it is rational for an agent to do *P.* One of the founders of expected utility theory, Savage (1954), did commit the so-called Allais paradox, thus disproving his own theory; but, when this was pointed out to him, he said that, on second thought, the normative appeal of expected utility theory was stronger and that he would now revise his choice accordingly. In another example, economists were very surprised that their intuition about behavior in so-called ultimatum games did not fit people's behavior (Roth 1995), even though non-economists would have thought the modal choice as the intuitive one. Because of their training, decision scientists may find it all too easy to consider, by introspection, their normative theories as having descriptive value for "normal" decision-makers, but that may only be because their intuition has

https://doi.org/10.1017/S0140525X04340069 Published online by Cambridge University Press

been honed and changed by years of training. Introspection can thus be overused to become a comfortable illusion for the theorist, and an obstacle for science. Experiments with oneself as subject may be easier in the decision sciences as implied by Roberts, but they are also of more limited value.

Author's Response

Self-experimentation: Friend or foe?

Seth Roberts

Department of Psychology, University of California at Berkeley, Berkeley, CA 94720-1650. **roberts@socrates.berkeley.edu**

Abstract: The topics discussed in this response are in four broad areas: (1) *Idea generation,* including the failure to discuss and teach idea generation and how to nurture new ideas (sect. R2), sources of ideas worth testing with self-experimentation (sect. R3), and unusual features of the situation that may have increased the discovery rate (sect. R4); (2) *Miscellaneous methodological issues,* such as the value of mental experiments (sect. R5) and the limitations of double-blind experiments (sect. R6); (3) *Subject-matter issues,* including the relationship of this work to other evolutionary psychology studies and a new way to test evolutionary explanations (sect. R7), as well as questions about the mood results (sect. R8) and the weight results (sect. R9); and (4) *Self-experimentation,* including its difficulty (sect. R10) and future (sect. R11).

R1. Introduction

Because self-experimentation is unfamiliar, commentators on the target article resemble early adopters, the first to consider buying new products, in their willingness to take it seriously. I am grateful for their thoughtful remarks.

R2. Missing traditions

The target article begins by noting the absence of data-gathering methods designed to generate ideas (see "Missing methods") but **Glenn** and **Cabanac** point out that this is part of a larger gap. Glenn saw in the target article "a sometimes subtle but pervasive message that our understanding of scientific method is limited by the grip of traditional formulations." One of those traditions is not discussing the origin of ideas. Glenn notes how, when she was a graduate student, discussions of scientific method began *after* the hypothesis had been formulated. This state of affairs has persisted; there is little discussion of idea generation in recent textbooks (e.g., Cozby 2004; Stangor 2004) and scientific articles that introduce a new idea rarely explain its source (Feynman 1972). Moreover, as Cabanac says, granting agencies have no place in their budgets for idea generation. Their attitude seems to be: *If it happens, fine, but don't expect us to pay for it.* This tradition, as Glenn says, creates missed opportunities – discoveries that could have been made but were not.

A closely related tradition is no discussion of how to nurture new ideas. If you have a new idea, what should you do? Test it, sure, but how? The target-article research made me vividly aware that I could not answer this question nor find anything written about it. Trial and error led to one answer: *Take baby steps,* that is, do the simplest, easiest test that could increase or decrease the plausibility of the new idea (Roberts 2001). In other words, raise the bar slowly. The larger the step – the larger the difference between what you plan to do and what you have already done – the more untested assumptions you are making. The more untested assumptions, the more likely one of them is wrong. This might seem obvious, but most research proposals I encounter involving new ideas do not abide by it. The result, as **Glenn** says, is a lot of wasted effort. A few years ago, I attended a research meeting at a leading school of public health. The eight or ten persons at the meeting (almost all professors or persons with Ph.D.s) were deciding how to do a study of asthma. They were about to begin a large and expensive experiment, involving about 100 families, to measure the effect of an apartment cleaning on the incidence of asthma and allergies. They had done a pilot study with two or three families; the next step was the study itself. I suggested that they first do a larger pilot study, pointing out that the existing pilot study had not tested several assumptions that the larger study was based on. No one agreed with this suggestion, and two of the persons with doctorates said they disagreed with it. The study went ahead with no additional pilot work and was a disaster. Recruitment of families turned out to be far more difficult than expected. Funding had come from the National Institutes of Health, supporting **Cabanac**'s doubts about the wisdom of granting agencies.

R3. Sour\ces of ideas to test

The target article might have been titled "Self-experimentation as a source of ideas for conventional research" or "Self-experimentation as a growth medium for new ideas," because, in each example, self-experimentation made a new idea more plausible. **Lubart & Mouchiroud** wonder where the nurtured ideas came from. "Being able to test ideas cheaply [via self-experimentation] allows one to entertain a wider set of ideas," they write. "But these ideas need to be generated."

How were these ideas generated? For each idea that led to or was supported by self-experimentation, Table R1 gives the precipitating event (what caused me to think of it) and states whether written records helped generate it. The line between *experiment* (done to see what happens) and *natural experiment* (a change that happens for other reasons) is blurry, because actions may have more than one goal. I ate less-processed food to lose weight (Example 1), but I also did it to find out if what I had been telling my students, based on research with rats, was true for humans. Table R1 calls it a natural experiment, because my main goal was to lose weight. Table R1 distinguishes between *short* experiments (lasting a few days or less) and *long* experiments (lasting longer), because a short experiment may be nothing more than trying something new.

Table R1 supports several commentators' views about how to generate ideas worth testing. **Schegloff** believes that "sustained and systematic observation" is under-appreciated. Indeed, Table R1 shows that written records were helpful in three or four cases. **Lubart & Mouchi-**

https://doi.org/10.1017/S0140525X04340069 Published online by Cambridge University Press