BEHAVIORAL AND BRAIN SCIENCES (2001) **24,** 383–451
Printed in the United States of America

Experimental practices in economics: A methodological challenge for psychologists?

Ralph Hertwig
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany.
hertwig@mpib-berlin.mpg.de

Andreas Ortmann
Center for Economic Research and Graduate Education, Charles University, and Economics Institute, Academy of Sciences of the Czech Republic, 111 21 Prague 1, Czech Republic.
andreas.ortmann@cerge.cuni.cz

Abstract: This target article is concerned with the implications of the surprisingly different experimental practices in economics and in areas of psychology relevant to both economists and psychologists, such as behavioral decision making. We consider four features of experimentation in economics, namely, script enactment, repeated trials, performance-based monetary payments, and the proscription against deception, and compare them to experimental practices in psychology, primarily in the area of behavioral decision making. Whereas economists bring a precisely defined "script" to experiments for participants to enact, psychologists often do not provide such a script, leaving participants to infer what choices the situation affords. By often using repeated experimental trials, economists allow participants to learn about the task and the environment; psychologists typically do not. Economists generally pay participants on the basis of clearly defined performance criteria; psychologists usually pay a flat fee or grant a fixed amount of course credit. Economists virtually never deceive participants; psychologists, especially in some areas of inquiry, often do. We argue that experimental standards in economics are regulatory in that they allow for little variation between the experimental practices of individual researchers. The experimental standards in psychology, by contrast, are comparatively laissez-faire. We believe that the wider range of experimental practices in psychology reflects a lack of procedural regularity that may contribute to the variability of empirical findings in the research fields under consideration. We conclude with a call for more research on the consequences of methodological preferences, such as the use on monetary payments, and propose a "do-it-both-ways" rule regarding the enactment of scripts, repetition of trials, and performance-based monetary payments. We also argue, on pragmatic grounds, that the default practice should be not to deceive participants.

Keywords: behavioral decision making; cognitive illusions; deception; experimental design; experimental economics; experimental practices; financial incentives; learning; role playing

1. Introduction

Empirical tests of theories depend crucially on the methodological decisions researchers make in designing and implementing the test (Duhem 1953; Quine 1953). Analyzing and changing specific methodological practices, however, can be a challenge. In psychology, for instance, "it is remarkable that despite two decades of counterrevolutionary attacks, the mystifying doctrine of null hypothesis testing is still today the Bible from which our future research generation is taught" (Gigerenzer & Murray 1987, p. 27). Why is it so difficult to change scientists' practices? One answer is that our methodological habits, rituals, and perhaps even quasi-religious attitudes about good experimentation are deeply entrenched in our daily routines as scientists, and hence often not reflected upon.

To put our practices into perspective and reflect on the costs and benefits associated with them, it is useful to look at methodological practices across time or across disciplines. Adopting mostly the latter perspective, in this arti-

RALPH HERTWIG is a research scientist at the Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development in Berlin. Currently he is a visiting scholar at Columbia University, New York. His research focuses on how people reason and make decisions when faced with uncertainty, the role of simple heuristics in human judgment and decision making, and how heuristics are adapted to the ecological structure of particular decision environments. In 1996, the German Psychological Association awarded him the Young Scientist Prize for his doctoral thesis.

ANDREAS ORTMANN is an assistant professor at the Center for Economic Research and Graduate Education at Charles University and researcher at the Economics Institute of the Academy of Sciences of the Czech Republic, both in Prague, and also a visiting research scientist at the Max Planck Institute for Human Development in Berlin. An economist by training, his game-theoretic and experimental work addresses the origins and evolution of languages, moral sentiments, conventions, and organizations.

© 2001 Cambridge University Press 0140-525X/01 $12.50

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

cle we point out that two related disciplines, experimental economics and *corresponding* areas in psychology (in particular, behavioral decision making) have very different conceptions of good experimentation.

We discuss the different conceptions of good experimentation in terms of four key variables of experimental design and show how these variables tend to be realized differently in the two disciplines. In addition, we show that experimental standards in economics, such as performance-based monetary payments (henceforth, *financial incentives*) and the proscription against deception, are rigorously enforced through conventions or third parties. As a result, these standards allow for little variation in the experimental practices of individual researchers. The experimental standards in psychology, by contrast, are comparatively laissez-faire, allowing for a wider range of practices. The lack of procedural regularity and the imprecisely specified social situation "experiment" that results may help to explain why in the "muddy vineyards" (Rosenthal 1990, p. 775) of soft psychology, empirical results "seem ephemeral and unreplicable" (p. 775).

1.1. The uncertain meaning of the social situation "experiment"

In his book on the historical origins of psychological experimentation, Danziger (1990) concluded that "until relatively recently the total blindness of psychological investigators to the social features of their investigative situations constituted one of the most characteristic features of their research practice" (p. 8). This is deplorable because the experimenter and the human data source are necessarily engaged in a social relationship; therefore, experimental results in psychology will always be codetermined by the social relationship between experimenter and participant. Schultz (1969) observed that this relationship "has some of the characteristics of a superior-subordinate one. . . . Perhaps the only other such one-sided relationships are those of parent and child, physician and patient, or drill sergeant and trainee" (p. 221). The asymmetry of this relationship is compounded by the fact that the experimenter knows the practices of experimentation by virtue of training and experience, while the typical subject is participating in any given experiment for the first time.[1]

Under these circumstances, and without clear-cut instructions from the experimenter, participants may generate a variety of interpretations of the experimental situation and therefore react in diverse ways to the experimental stimuli. In the words of Dawes (1996):

> The objects of study in our experiments (i.e., people) have desires, goals, presuppositions, and beliefs about what it is we wish to find out. Only when it is explicitly clear that what we are seeking is maximal performance . . . can we even safely assume that our interpretation of the experimental situation corresponds to that of our subjects. . . . Even then, however, we may not be able to . . . "control for" factors that are not those we are investigating. (p. 20)

1.2. Defining the social situation "experiment"

In this article, we argue that experimental standards in economics reduce participants' uncertainty because they require experimenters to specify precisely the "game or contest" (Rieken 1962, p. 31) between experimenter and participant in a number of ways. In what follows, we consider four key features of experimental practices in economics, namely, script enactment, repeated trials, financial incentives, and the proscription against deception. The differences between psychology and economics on these four features can be summed up – albeit in a simplified way – as follows. Whereas economists bring a precisely defined "script" to experiments and have participants enact it, psychologists often do not provide such a script. Economists often repeat experimental trials; psychologists typically do not. Economists almost always pay participants according to clearly defined performance criteria; psychologists usually pay a flat fee or grant a fixed amount of course credit. Economists do not deceive participants; psychologists, particularly in social psychology, often do.

We argue that economists' realizations of these variables of experimental design reduce participants' uncertainty by explicitly stating action choices (script), allowing participants to gain experience with the situation (repeated trials), making clear that the goal is to perform as well as they can (financial incentives), and limiting second-guessing about the purpose of the experiment (no deception). In contrast, psychologists' realizations of these variables tend to allow more room for uncertainty by leaving it unclear what the action choices are (no script), affording little opportunity for learning (no repeated trials), leaving it unclear what the experimenters want (no financial incentives), and prompting participants to second-guess (deception).

Before we explore these differences in detail, four caveats are in order. First, the four variables of experimental design we discuss are, in our view, particularly important design variables. This does not mean that we consider others to be irrelevant. For example, we question economists' usual assumption that the abstract laboratory environment in their experiments is neutral and, drawing on results from cognitive psychology, have argued this point elsewhere (Ortmann & Gigerenzer 1997). Second, we stress that whenever we speak of standard experimental practices in "psychology," we mean those used in research on behavioral decision making (an area relevant to both psychologists and economists; e.g., Rabin 1998) and related research areas in social and cognitive psychology such as social cognition, problem solving, and reasoning. The practices discussed and the criticisms leveled here do not apply (or do so to a lesser degree), for instance, to research practices in sensation and perception, biological psychology, psycho-physics, learning, and related fields. Third, we do not provide an exhaustive review of the relevant literature, which, given the wide scope of the paper, would have been a life's work. Rather, we use examples and analyze several random samples of studies to show how differences in the way the design variables are realized can affect the results obtained. Moreover, even in discussing the limited areas of research considered here, we are contrasting prototypes of experimental practices to which we are aware many exceptions exist.

Finally, we do not believe that the conventions and practices of experimental economists constitute the gold standard of experimentation. For example, we concur with some authors' claim that economists' strict convention of providing financial incentives may be too rigid and may merit reevaluation (e.g., Camerer & Hogarth 1999). The case for such reevaluation has also been made in a recent symposium in *The Economic Journal* (e.g., Loewenstein

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

1999). This symposium also takes issue with the assumed neutrality of the laboratory environment (e.g., Loomes 1999), scripts that are too detailed (e.g., Binmore 1999; Loewenstein 1999; Loomes 1999; Starmer 1999), and the relevance of one-shot decision making (e.g., Binmore 1999; Loewenstein 1999), among other aspects of experimentation in economics that warrant reevaluation (e.g., Ortmann & Tichy 1999). In other words, a paper entitled "Experimental practices in psychology: A challenge for economists?" may well be worth writing.

2. Enacting a script versus "ad-libbing"

Economists run experiments usually for one of three reasons: to test decision-theoretic or game-theoretic models, to explore the impact of institutional details and procedures, or to improve understanding of policy problems such as the behavior of different pricing institutions (e.g., Davis & Holt 1993, Ch. 3 on auctions and Ch. 4 on posted offers).

To further understanding of policy problems, experimental economists construct small-scale abstractions of real-world problems (although typically these miniature replicas are framed in abstract terms). To test theoretical models, economists attempt to translate the model under consideration into a laboratory set-up that is meant to capture the essence of the relevant theory. This mapping inevitably requires the experimenter to make decisions about "institutional details" (i.e., the degree of information provided in the instructions, the way the information is presented to participants, the communication allowed between participants, etc.). Economists have learned to appreciate the importance of such institutional details and procedures, and how these might affect results (e.g., Davis & Holt 1993, pp. 507–509; Osborne & Rubinstein 1990; Zwick et al. 1999).

To enhance replicability and to trace the sometimes subtle influence of institutional details and experimental parameters, experimental economists have come to provide participants with scripts (instructions) that supply descriptions of players, their action choices, and the possible payoffs (for standard examples of such instructions, see appendices in Davis & Holt 1993). Economists then ask participants to enact those scripts. For example, they assign each of them the role of buyer or seller and ask them to make decisions (e.g., to buy or sell assets) that determine the amount they are paid for their participation, a practice discussed in detail later.

An example of a script and its enactment is provided by Camerer et al. (1989) in their investigation of hindsight bias.[2] In their design, an "uninformed" group of participants guessed future earnings of real companies based on information such as the previous annual earnings per share. An "informed" group of participants (who were told the actual earnings) then traded assets that paid dividends equal to the earnings predicted by the uninformed group. Participants in both groups were provided with a precise script. Those in the uninformed group were given the role (script) of a market analyst faced with the task of predicting the future dividends of various companies. Those in the informed group were assigned the role of trader: they knew that the dividend was determined by the uninformed group's predictions. Thus, to price the assets optimally (and thereby to avoid hindsight bias), the "traders" had to predict the prediction of the "analysts" accurately, that is, to ignore their knowledge of the actual dividends. Eventually, the traders traded the assets to others in actual double-oral auctions, in which "buyers and sellers shouted out bids or offers at which they were willing to buy or sell. When a bid and offer matched, a trade took place" (p. 1236).

Unlike Camerer et al.'s (1989) study, typical hindsight bias experiments in psychology do not provide participants with a script, thus forcing them to ad-lib, that is, to infer the meaning of the experiment as they go. In a typical study (Davies 1992), participants were given a series of assertions and asked to rate the truth of each. They were then given feedback (i.e., the truth values of the assertions) and later asked to recall their original judgment. In contrast to Camerer et al. (1989), Davies did not assign specific roles to participants or provide them with any precise script. Instead, the first stage of the study, during which participants rated assertions for their truth, was merely described to participants as "involving evaluation of college students' knowledge" (Davies 1992, p. 61), and they were told that the recollection stage "concerned people's ability to remember or recreate a previous state of knowledge" (Davies 1992, p. 61). This procedure is typical of many psychological studies on hindsight bias (e.g., Hell et al. 1988; Hoffrage & Hertwig 1999).

In psychological research on judgment, decision making, and reasoning, too, researchers typically do not provide participants with a script to enact. Much of this research involves word problems such as the conjunction task (e.g., Tversky & Kahneman 1983), the engineer-lawyer task (e.g., Kahneman & Tversky 1973), the Wason selection task (e.g., Evans et al. 1993), and the 2-4-6 task (e.g., Butera et al. 1996). These problems share a number of typical features. For example, they often are ambiguous (e.g., use polysemous terms such as "probability," see Hertwig & Gigerenzer 1999) and require participants to ignore conversational maxims in order to reach the "correct" solution (see Hilton 1995).[3] Furthermore, they do not require participants to assume clearly specified roles, like the analysts and traders in Camerer et al.'s (1989) study, or to enact a script. As a result, participants are forced to ad-lib.

Participants' ad-libbing is likely to be influenced by their expectations about what experimenters are looking for. Providing a script would not alter the fact that the typical participant in psychology (and economics) has never or rarely encountered a particular experimental situation before. That is, notwithstanding provision of a script, participants are still likely to be sensitive to cues that are communicated to them by means of campus scuttlebutt, the experimenter's behavior, and the research setting. However, scripts can constrain participants' interpretations of the situation by focusing their attention on those cues that are intentionally communicated by the experimenter (e.g., the task instructions), thus clarifying the demand characteristics of the social situation "experiment." As a consequence, scripts may enhance replicability.

Enacting a script is closely related to "role playing" in social psychology (e.g., Greenwood 1983; Krupat 1977), in which the "intent is for the subject to directly and actively involve himself in the experiment, and to conscientiously participate in the experimental task" (Schultz 1969, p. 226). To borrow the terminology of Hamilton's useful three-dimensional classification (referred to in Geller 1978, p. 221), the role-playing simulations that come closest to econom-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ics experiments are those performed (rather than imagined) and scripted (rather than improvised), and in which the dependent variable is behavior (rather than verbal utterances). In economics experiments, however, participants do not just simulate but are real agents whose choices have tangible consequences for them. For example, in the Camerer et al. (1989) study, they were real analysts and real traders, albeit in a scaled-down version of a real market.

2.1. Does providing and enacting a script matter?

We believe that providing a script for participants to enact affects experimental results. At the same time, we readily admit that the evidence for this claim is at present tenuous because provision of scripts and their enactment are rarely treated as independent variables. Using as examples the prediction task in Camerer et al.'s (1989) study and the Wason selection task in psychology, we now discuss the potential importance of providing a script and having participants enact it.

Camerer et al. (1989) compared the amount of hindsight bias in the predictions of participants who enacted the role of trader (i.e., who actually traded assets in the double-oral auction) to the bias in predictions made by another group of participants who did not enact the role of trader. The goal of the two groups was the same: to predict the average prediction of the uninformed group, given companies' actual earnings. Both groups received incentives for making correct predictions. Camerer et al. (1989) reported that participants in both conditions exhibited some hindsight bias, but enactment of the trader role reduced the bias by about half: The difference in hindsight bias between the two groups was $r = .18$ (calculated from data in their Fig. 4), a small to medium effect (see Rosenthal & Rosnow 1991, p. 444).

Research on the Wason selection task provides another example of a situation in which providing a script (or more precisely, a proxy for one) – namely, assigning participants to the perspective of a particular character – dramatically changes their responses. This task is perhaps the most studied word problem in cognitive psychology. In what is known as its abstract form, participants are shown four cards displaying symbols such as *T*, *J*, *4*, and *8* and are given a conditional rule about the cards, such as "If there is a *T* on one side of the card [antecedent *P*], then there is a *4* on the other side of the card [consequent *Q*]." Participants are told that each card has a letter on one side and a number on the other. They are then asked which cards they would need to turn over in order to discover whether the conditional rule is true or false. The typical result, which has been replicated many times (for a review, see Evans et al. 1993, Ch. 4), is that very few participants (typically only about 10%) give the answer prescribed by propositional logic: *T* and *8* (*P* & *not-Q*). Most participants choose either *T* (*P*) alone or *T* and *4* (*P* & *Q*). These "errors" in logical reasoning have been seen as reflections of the confirmation bias, the matching bias, and the availability heuristic (for a review, see Garnham & Oakhill 1994).

The original, abstract Wason selection task was content-free. Numerous researchers have since shown that dressing it in thematic garb, that is, putting it in a social context, increases the percentage of logically correct answers. In one such task, a police officer is checking whether people conform to certain rules: in the context of a drinking age law ("If someone is drinking beer [*P*], then they must be over 19 years of age [*Q*]"), 74% of participants gave the *logical P & not-Q* response (Griggs & Cox 1982). Gigerenzer and Hug (1992) later demonstrated that the way in which social context affects reasoning in the selection task also depends on the *perspective* into which participants are cued. For instance, the implications of the rule "If an employee works on the weekend, then that person gets a day off during the week" depend on whether it is seen from the perspective of an employer or of an employee. Among participants cued into the role of an employee, the dominant answer was *P & not-Q* (75%); among participants cued into the role of an employer, in contrast, the dominant response was *not-P & Q* (61%; for more detail, see Ortmann & Gigerenzer 1997). Perspective can thus induce people to assume certain social roles, activating a script like those provided in economics experiments.[4]

To conclude this section, the effects of role playing in Camerer et al.'s (1989) study and perspective taking in selection tasks suggest that supplying a script for participants to enact can make an important difference to the results obtained. Although script provision (i.e., action choices, payoffs, perspective, etc.) demands more elaborate and transparent instructions (e.g., compare Camerer et al.'s market study with any typical hindsight bias study in psychology), it is likely to reduce the ambiguity of the experimental situation and thereby increase researchers' control over participants' possible interpretations of it. This practice is also likely to enhance the replicability of experimental results. We propose that psychologists consider having participants enact scripts wherever possible.

3. Repeated trials versus snapshot studies

Economists use repeated trials for (at least) two reasons. The first is to give participants a chance to adapt to the environment, that is, to accrue experience with the experimental setting and procedure. This motivation applies to both decision and game situations and reflects economists' interest in the impact of experience on behavior. Binmore (1994) articulated this rationale as follows:

> But how much attention should we pay to experiments that tell us how inexperienced people behave when placed in situations with which they are unfamiliar, and in which the incentives for thinking things through carefully are negligible or absent altogether? . . . Does it [the participant's behavior] survive after the subjects have had a long time to familiarize themselves with all the wrinkles of the unusual situation in which the experimenter has placed them? If not, then the experimenter has probably done no more than inadvertently trigger a response in the subjects that is adapted to some real-life situation, but which bears only a superficial resemblance to the problem the subjects are really facing in the laboratory. (pp. 184–85)

The second motivation for the use of repeated trials, while also reflecting economists' interests in the impact of experience on behavior, is specific to game situations. Repeated trials afford participants the opportunity to learn how their own choices interact with those of other players in that specific situation. Although in practice the two kinds of learning are difficult to distinguish, they are conceptually distinct. The first kind of learning (adapting to the laboratory environment) relates to a methodological concern that participants may not initially understand the laboratory environment and task, whereas the second kind of learning

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

(understanding how one's own choices interact with those of other participants) relates to the understanding of the possibly strategic aspects of the decision situation. Game theory captures those strategic aspects and suggests that for certain classes of games, people's behavior "today" will depend on whether and how often they may be paired with others in the future.

Underlying both motivations for the use of repeated trials is economists' theoretical interest in equilibrium solutions, that is, the hope that for every scenario a belief or behavior exists that participants have no incentive to change. However, equilibrium is assumed not to be reached right away. Rather, it is expected to evolve until participants believe their behavior to be optimal for the situation they have been placed in. This is why in economics experiments "special attention is paid to the last periods of the experiment . . . or to the change in behavior across trials. Rarely is rejection of a theory using first-round data given much significance" (Camerer 1997, p. 319). Note, however, that although economists tend to use repeated trials most of the time, there are important exceptions. For instance, most studies of trust games (e.g., Berg et al. 1995), dictator games (e.g., Hoffman et al. 1996), and ultimatum games employ one-shot scenarios. It is interesting to consider whether the attention-grabbing results of these games are due to the very fact that they are typically implemented as one-shot rather than repeated games.

Typically, economists implement repeated trials either as *stationary replications* of one-shot decision and game situations or as repeated game situations. Stationary replication of simple decision situations (i.e., without other participants) involves having participants make decisions repeatedly in the same one-shot situation. Stationary replication of game situations also involves having participants make decisions repeatedly in the same one-shot situation, but with new participants in each round. In contrast, other repeated game situations may match participants repeatedly with one another and thus allow for strategic behavior. Neither stationary replication of one-shot decision and game situations nor other repeated game situations implement environments that change. Instead, learning is typically studied in environments whose parameterization (e.g., payoff structure) does not change. Camerer (1997) referred to such situations as "'Groundhog Day' replication" (p. 319). In what follows, we focus on the special case of Groundhog Day replication referred to as stationary replication above.

In contrast to economists, researchers in behavioral decision making typically provide little or "no opportunity for learning" (Thaler 1987, p. 119; see also Hogarth 1981; Winkler & Murphy 1973), tending instead to conduct "snapshot" studies. It would be misleading, however, to suggest that psychologists have ignored the role of feedback and learning. For instance, there is a history of multi-stage decision making in research on behavioral decision making (see Rapoport & Wallsten 1972). Moreover, studies in which repetition and feedback are used can be found in research on multiple-cue probability learning (e.g., Balzer et al. 1989; Klayman 1988), social judgment theory (Hammond et al. 1975), dynamic decision making (e.g., Brehmer 1992; 1996; Diehl & Sterman 1995; Edwards 1962), probabilistic information processing (e.g., Wallsten 1976), and in research on the effects of different kinds of feedback (e.g., Creyer et al. 1990; Hogarth et al. 1991). Nevertheless, "most judgment research has focused on discrete events. This has led to underestimating the importance of feedback in ongoing processes" (Hogarth 1981, p. 197).

To quantify the use of repeated trials and feedback in behavioral decision making, we analyzed a classic area of research in this field, namely, that on the base-rate fallacy. For the last 30 years, much research has been devoted to the observation of "fallacies," "biases," or "cognitive illusions" in inductive reasoning (e.g., systematic deviations from the laws of probability). Among them, the base-rate fallacy "had a celebrity status in the literature" (Koehler 1996, p. 2). According to Koehler's (1996) recent review of base-rate fallacy research, "hundreds of laboratory studies have been conducted on the use of base rates in probability judgment tasks" (p. 2), and "investigators frequently conclude that base rates are universally ignored" (p. 2). How many of these laboratory studies have paid attention to the possible effects of feedback and learning?

To answer this question, we examined the articles cited in Koehler's (1996) comprehensive review of Bayesian reasoning research. We included in our analysis all empirical studies on the use of base rates published in psychology journals (excluding journals from other disciplines and publications other than articles) since 1973, the year in which Kahneman and Tversky published their classic study on the base-rate fallacy. This sample comprises a variety of paradigms, including, for instance, word problems (e.g., engineer-lawyer and cab problems), variations thereof, and "social judgment" studies (which explore the use of base rates in social cognition such as stereotype-related trait judgments). As the unit of analysis, we took studies – most articles report more than one – in which an original empirical investigation was reported.

By these criteria, 106 studies were included in the analysis. Although this sample is not comprehensive, we believe it is representative of the population of studies on the use of base rates. Of the 106 studies, only 11 (10%) provided participants with some kind of trial-by-trial feedback on their performance (Study 1 in Manis et al. 1980; Studies 1 and 2 in Lopes 1987; Lindeman et al. 1988; Studies 1–5 in Medin & Edelson 1988; Studies 1 and 2 in Medin & Bettger 1991). The picture becomes even more extreme if one considers only those studies that used (sometimes among others) the classic word problems (engineer-lawyer and cab problem) employed by Kahneman and Tversky (1973) or variants thereof. Among these 36 studies, only 1 provided trial-by-trial feedback concerning participants' posterior probability estimates (Lindeman et al. 1988). Based on this survey, we conclude that repetition and trial-by-trial feedback is the exception in research on the base-rate fallacy. This conclusion is consistent with that drawn by Hogarth (1981) almost 20 years ago, namely, that "many discrete judgment tasks studied in the literature take place in environments degraded by the lack of feedback and redundancy. . . . As examples, consider studies of Bayesian probability revision" (p. 199).

3.1. Do repetition and feedback matter?

There is evidence from economists' research on the use of base rates involving repeated trials that they do, indeed. When trials are repeated, base rates do not seem to be universally ignored. Harrison (1994) designed an experiment to test, among other things, the effect of repetition (plus feedback) and the validity of the representativeness heuris-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

tic, which Kahneman and Tversky (1973) proposed as an explanation for people's "neglect" of base rates. This explanation essentially states that people will judge the probability of a sample by assessing "the degree of correspondence [or similarity] between a sample and a population" (Tversky & Kahneman 1983, p. 295).

Unlike Kahneman and Tversky (1973), but like Grether (1980; 1992), Harrison used a bookbag-and-poker-chips paradigm in which participants had to decide from which of two urns, A and B, a sample of six balls (marked with either Ns or Gs) had been drawn. In addition to the ratio of Ns and Gs in the sample and the frequencies of Ns and Gs in the urns (urn A: four Ns and two Gs, urn B: three Ns and three Gs), participants knew the urns' priors (i.e., the probabilities with which each of the two urns was selected). In this design, the ratio of Ns and Gs in the sample can be chosen so that use of the representativeness heuristic leads to the choice of urn A (as the origin of the sample of six balls), whereas application of Bayes's theorem leads to the choice of urn B, and vice versa.

Participants in Harrison's (1994) study judged a total of 20 samples. After each one, participants were told from which urn the balls were drawn. After each set of 10 decisions, their earnings were tallied based on the number of choices they made in accordance with Bayes's theorem. There were three payoff schedules: Two were contingent on performance and one was not. Harrison (1994) split the choices according to whether they were made when participants were "inexperienced" (first set of 10 decisions) or "experienced" (second set of 10 decisions). He found that the representativeness heuristic strongly influenced the decisions of participants who were inexperienced and unmotivated, that is, who had completed only the first set of 10 decisions and who received a fixed amount of money (independent of performance). However, he also found that when those participants who were not monetarily motivated made the *second* set of 10 decisions, "the Heuristic has no noticeable influence at all" (pp. 249–50). Moreover, Harrison (1994) reported finding little to no evidence of the representativeness heuristic among inexperienced participants (i.e., in the first set of 10 decisions) whose earnings were based on performance.

Harrison's (1994) results seem to contradict Grether's (1980). Grether concluded that participants do tend to follow the representativeness heuristic. However, Grether employed a different definition of experience. Specifically, he counted every participant who had previously assessed the *same* prior-sample combination as experienced. In Harrison's study, in contrast, participants had to make 10 judgments with feedback before they were considered experienced. That experience can substantially improve Bayesian reasoning has also been shown in a series of experiments by Camerer (1990); he also observed that the significance of the biases increased because the variance decreased with experience. The three studies taken together strongly suggest that one ought to use repeated trials when studying Bayesian reasoning, and that biases diminish in magnitude with sufficient experience (Camerer 1990; Harrison 1994), although not necessarily after only a few trials (Grether 1980).

This conclusion is also confirmed by a set of experiments conducted in psychology. In Wallsten's (1976) experiments on Bayesian revision of opinion, participants completed a large number of trials. In each trial, participants observed events (samples of numbers), decided which of two binomial distributions was the source, and estimated their confidence in the decision. Participants received trial-by-trial feedback, and the sampling probabilities of the two populations under consideration changed from trial to trial. The results showed strong effects of experience on Bayesian reasoning. In the early trials, participants tended to ignore the sampling probability under the less likely hypothesis. As they gained experience, however, they increasingly gave more equal weight to the likelihood of the data under each of the two hypotheses (also see Wallsten 1972).

What are the results in the few studies in our sample that examined the use of base rates using trial-by-trial feedback? Only 4 of these 11 studies (Lindeman et al. 1988; Lopes 1987; Manis et al. 1980) systematically explored the effect of repetition and feedback by comparing a feedback and a no-feedback condition. Table 1 summarizes the results of these four studies. Although the small sample size limits the generalizability of the findings, the results in Table 1 indicate that providing people with an opportunity to learn does increase the extent to which base rates are used, and does bring Bayesian inferences closer to the norm.

However, cautionary notes are in order: Manis et al.'s findings have been suggested to be consistent with reliance on representativeness (Bar-Hillel & Fischhoff 1981); in Lindeman et al.'s (1988) study, the effect of learning did not generalize to a new problem (which according to Lindeman et al. could be due to a floor effect), and in Lopes's (1987) studies the effects of performance-dependent feedback and a training procedure cannot be separated. More generally, Medin and Edelson (1988) caution that people's use of base-rate information "must be qualified in terms of particular learning strategies, category structures, and types of tests" (p. 81).

In the same sample of studies, we also found some that investigated the effect on Bayesian reasoning of "mere practice," that is, the use of repeated trials without feedback. According to these studies, even mere practice can make a difference. With repeated exposure, it seems that "respondents tended to be influenced by the base rate information to a greater degree" (Hinsz et al. 1988, p. 135; see also Fischhoff et al. 1979, p. 347). Moreover, mere practice seems to increase slightly the proportion of Bayesian responses (Gigerenzer & Hoffrage 1995), and can increase markedly participants' consistency (i.e., in applying the same cognitive algorithm across tasks). Mere practice also may drastically alter the distribution of responses: "in the one-judgment task, subjects appear to respond with one of the values given, whereas when given many problems, they appear to integrate the information" (Birnbaum & Mellers, 1983 p. 796).

Taken together, these examples illustrate that repetition of trials combined with performance feedback, and to some extent even mere practice (repetition without feedback), can improve participants' judgments in tasks in which it has been alleged that "information about base rates is generally observed to be ignored" (Evans & Bradshaw 1986, p. 16).

Research on the base-rate fallacy is not the only line of research in behavioral decision making where feedback and repetition seem to matter. Another example is research on "preference reversals," in which most participants choose gamble A over B but then state that their minimum willingness-to-accept price for A is less than the price of B

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Table 1. *Effects of trial-by-trial feedback on the use of base rates obtained in studies from Koehler's (1996) review*

	Task/Response/Feedback	Comparison	General results
Manis, Dovalina, Avis & Cardoze (1980)	*Task.* In each of 50 trials, participants viewed a photograph of a young male, guessed his attitude on an issue such as legalization of the use of marijuana. *Response.* Binary classification (pro vs. con). *Feedback.* After each trial, feedback on binary classification.	Study 1 *Two feedback conditions.* 80% vs. 20% "pro" attitude base rates *Control condition.* No feedback	"Base-rate information had a clear-cut effect on predictive behavior" (p. 235). By the end of the 50 trials, respondents in the no-feedback condition predicted that about half of the pictured men have a "pro" attitude, whereas in the two feedback conditions, the average percentage of predicted "pro" attitude was 70% and 30% in the 80% and 20% base-rate conditions, respectively.
Lopes (1987)	*Task.* In each of 175 trials, participants were required to decide from which of two alternative Bernoulli processes a sample was generated.[a] *Response.* Rating as to whether the sample comes from process 1 or 2. *Feedback.* On 13 practice and training trials, participants received training and feedback on a judgment operation.	Study 1 *Training condition* *Control condition* Received no training	Trained participants made fewer "adjustment errors" than untrained participants for "strong-weak" pairs (*eta* = .84) and diagonal pairs (*eta* = .42) but not for "weak-strong" pairs, and were more accurate in their judgments: Root-mean-squared deviations between obtained and optimal judgments were .07 and .02 for untrained participants, and .02 and .005 for trained participants, respectively. However, training did not make participants' responses *qualitatively* more Bayesian (p. 174).
	Same task but modified training (182 trials, of which 20 were practice and feedback trials)	Study 2 *Training condition* *Control condition*	"Clearly, the training procedure has been effective in bringing subjects' responses closer to optimal" (p. 178). Root-mean-squared deviations between obtained and optimal response were .05 and .10 for trained and untrained participants, respectively. The optimal response requires an interaction effect that accounts for 4.66% of the systematic sums of square (in terms of analysis of variance). In the training condition, this value is 4.55%, whereas in the control condition the value is .42%.
Lindeman, van den Brink & Hoogstraten (1988)	*Task.* In each of 16 trials, participants predicted the probability that a person described in a personality sketch has a certain profession (engineer-lawyer problem: two different base rates per description); after the training phase all participants were presented with a new problem (divorce problem) *Response.* Probability estimate *Feedback.* After the second estimate per description, feedback on the required probability estimate for second estimate	*Feedback Training-only* 32 probability estimates but no feedback *No treatment* No training	For the engineer-lawyer problem (the training problem), "over-all performance of subjects who received feedback was closer to the Bayesian norm than the performance of subjects who received training only" (p. 346). In the training-only condition, 50% of the mean differences between obtained and Bayesian probability estimates were significant (at p = .05); in the feedback condition, in contrast, only 13% of the mean differences were significant. Although the positive effects of feedback "do not generalize to" (p. 349) the divorce problem, thinking-aloud protocols show that base rates were less often mentioned in the no-treatment condition (n = 3 out of 16 porotcols, 19%) than in the feedback condition (n = 11 out of 18 protocols, 61%)[b].

[a]The context story asked participants to imagine making judgments concerning the maintenance of milling machines. The judgment concerns whether or not a crucial part has broken inside the machine. For each machine, participants received two samples; there were three kinds of pairs of samples (weak-strong, strong-weak, and diagonal; only the latter two are test cases for the training as here intuitive and normative response seem to diverge). In Study 1, the training procedure taught participants only that adjustment of the initial rating made after presentation of the second sample should always be in the direction of the hypothesis favored by the second sample. In Study 2, the training also taught participants to process the two samples in the order of their apparent relative strength.

[b]After the training phase, participants in all three conditions were presented with a different problem. Lindeman et al. (1988) speculated that the failure to find an effect of training for the estimates on this problem may be related to a floor effect: "the divorce problem is inherently easier than the training problem, so that subjects might get the test item right even without the training phase" (p. 349).

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

(Lichtenstein & Slovic 1971). This basic finding has been replicated many times with a great variety of gambles. In a repeated context, however, preference reversals are not as recalcitrant as this research makes them seem. For instance, Berg et al. (1985), Hamm (reported in Berg et al. 1985), and Chu and Chu (1990) observed that the number of preference reversals decreases if participants repeat the experiment. Berg et al. (1985) concluded that "these findings are consistent with the idea that economic theory describes the asymptotic behavior of individuals after they have become acclimated to the task" (p. 47). Chu and Chu (1990), who embedded their study in a market context, concluded that "*three transactions were all that was needed to wipe out preference reversals completely*" (p. 909, their emphasis).

Some have questioned the importance of learning (e.g., Brehmer 1980). Thaler (1987), among others, has argued that the equilibrium and convergence argument is misguided because "when major decisions are involved, most people get too few trials to receive much training" (p. 122). While it may be true that for some situations there is little opportunity for training, it is noteworthy that novices in real-life settings often have the opportunity to seek advice from others in high-stake "first trials" – an option not available in most experiments in both psychology and economics. Moreover, in the first trial, a novice might use a range of other strategies, such as trying to convert the task into hedge trimming rather than tree felling (Connolly 1988) in order to get feedback, holding back reserves, or finding ways to avoid firm commitments.

To conclude this section, testing a stimulus (e.g., a gamble, an inference task, a judgment task, or a choice task) only once is likely to produce high variability in the obtained data (e.g., less consistency in the cognitive processes). In the first trial, the participant might still be in the process of trying to understand the experimental instructions, the setting, the procedure, and the experimenter's intentions. The more often the participant works on the same stimulus, the more stable the stimulus interpretation (and the less pronounced the test anxiety; Beach & Phillips 1967) and the resulting behavior (as long as the situation is incentive-compatible and participants are neither bored nor distracted). People's performance in early trials, in other words, does not necessarily reflect their reasoning *competence* in later trials. We propose that psychologists consider using stationary replication, that is, repetition of one-shot decisions and game situations as well as feedback, and not restrict their attention to one-shot trials in which participants may be confused and have not had an opportunity to learn.

Last but not least, which design is appropriate is not only a methodological issue. The appropriateness of a design depends crucially on what aspects of behavior and cognition a given theory is designed to capture. Although recently economists have become increasingly interested in learning, prevailing theories in economics still focus on equilibrium behavior. In contrast, many (but not all) psychological judgment and decision-making theories are not explicit about the kind of behavior they target – first impressions, learning, or equilibrium behavior – and also do not explicate how feedback and learning may affect it. Clearly, if theories in psychology were more explicit about the target behavior, then the theories rather than the experimenter would define the appropriate test conditions, and thus questions about whether or not to use repeated trials would be less likely to arise.

4. Financial incentives versus no incentives

Although important objections have been raised to the way financial incentives are often structured (e.g., Harrison 1989; 1992), experimental economists who do not use them at all can count on not getting their results published. Camerer and Hogarth (1999) reported that "a search of the *American Economic Review* from 1970–1997 did not turn up a single published experimental study in which subjects were not paid according to performance" (p. 31). As Roth (1995) observed, "the question of actual versus hypothetical choices has become one of the fault lines that have come to distinguish experiments published in the economics journals from those published in psychology journals" (p. 86).

Economists use financial incentives for at least four reasons. The first is the widespread belief among experimental economists that salient payoffs (rewards or punishment) reduce performance variability (Davis & Holt 1993, p. 25). The second is the assumption that the saliency of financial incentives is easier to gauge and implement than most alternative incentives. The third is the assumption that most of us want more of it (so it is fairly reliable across participants), and there is no satiation over the course of an experiment (not so with German chocolate cake, grade points, etc.). The fourth, and arguably the most important argument motivating financial incentives, is that most economics experiments test economic theory, which provides a comparatively unified framework built on maximization assumptions (of utility, profit, revenue, etc.) and defines standards of optimal behavior. Thus, economic theory lends itself to straightforward translations into experiments employing financial incentives.

This framework is sometimes interpreted as exclusively focusing on the monetary structure at the expense of the social structure. We believe this to be a misunderstanding. Every experiment that employs financial incentives implicitly also suggests something about other motivators (e.g., altruism, trust, reciprocity, or fairness). For example, if in prisoner's dilemma games (or public good, trust, ultimatum, or dictator games) the behavior of participants does not correspond to the game-theoretic predictions, that is, if they show more altruism (trust, reciprocity, or fairness) than the theory predicts, then these findings also tell us something about the other nonmonetary motivators (assuming that demand effects are carefully controlled, and the experiments successfully implement the game-theoretic model).

Psychologists typically do not rely on a similarly unified theoretical framework that can be easily translated into experimental design. Moreover, in some important psychological domains, standards of optimal behavior are not as clearly defined (e.g., in mate choice), if they can be defined at all, or conflicting norms have been proposed (e.g., in hypothesis testing, probabilistic reasoning). In addition, there is the belief that "our subjects are the usual middle-class achievement-oriented people who wish to provide [maximal performance]" (Dawes 1996, p. 20), which seems to suggest that financial incentives are superfluous. Along similar lines, Camerer (1995) observed that "psychologists

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

presume subjects are cooperative and intrinsically motivated to perform well" (p. 599).

To quantify how different the conventions in economics and psychology are with regard to financial incentives, we examined all articles published in the *Journal of Behavioral Decision Making* (*JBDM*) in the 10-year period spanning 1988 (the year the journal was founded) to 1997. We chose *JBDM* because it is one of the major outlets for behavioral decision researchers and provides a reasonably representative sample of the experimental practices in this domain. As our unit of analysis we took experimental studies – a typical *JBDM* article reports several – in which some kind of performance criterion was used, or in which participants were provided with an explicit choice scenario involving monetary consequences.

In addition to studies in which no performance criterion was specified, we excluded studies in which no financial incentives could have been employed because experimenters compared performance across rather than within participants (i.e., between-subjects designs). In addition, we excluded studies in which the main focus was not on the performance criterion – either because it was only one among many explored variables or because processes rather than outcomes were examined. Finally, we omitted studies in which experimenters explicitly instructed participants that there were no right or wrong answers, or that we could not classify unequivocally (e.g., ambiguous performance criteria, or the description of the study leaves it open whether financial incentives were employed at all).

Our criteria were intentionally strict and committed us to evaluating each study in its own right and not with respect to some ideal study (e.g., we did not assume that each study that explored the understanding of verbal and numerical probabilities could have employed financial incentives only because Olson & Budescu, 1997, thought of an ingenious way to do it). These strict criteria stacked the deck against the claim that psychologists hardly use payments, as studies that could have employed payments if run differently were excluded.

We included 186 studies in the analysis. Out of those 186 studies, 48 (26%) employed financial incentives. Since *JBDM* publishes articles at the intersection of psychology, management sciences, and economics, and experimental economists such as John Hey and David Grether are on the editorial board, this ratio is very likely an overestimate of the use of financial incentives in related domains of psychological research. If one subtracts studies in which at least one of the authors is an economist or is affiliated with an economics department, then the percentage of studies using financial incentives declines to 22% (40 of 178 studies). If one additionally subtracts studies in which at least one of the authors is one of the few psychologists in behavioral decision making who frequently or exclusively use monetary incentives (Budescu, Herrnstein, Rapoport, and Wallsten), then the ratio declines still further to 15% (25 of 163). This survey suggests that financial incentives are indeed not the norm in behavioral decision making.

Our conclusion is also supported by a second sample of studies that we analyzed. As described in section 3, we examined 106 studies on Bayesian reasoning. These studies were published in a variety of journals, including journals from social psychology (e.g., *Journal of Personality and Social Psychology, Journal of Experimental Social Psychology*), cognitive psychology (e.g., *Cognition, Cognitive Psychology*), and judgment and decision making (e.g., *Organizational Behavior and Human Decision Processes, JBDM*). Thus, this sample represents a cross-section of journals. Of these 106 base-rate studies, only three provided financial incentives (Studies 1 and 2 in Nelson et al. 1990; and possibly Kahneman & Tversky's 1973 study).

4.1. Do financial incentives matter?

Given the typical economist's and psychologist's sharply diverging practices, it is not surprising to see diverging answers to the question of whether financial incentives matter. There is overwhelming consensus among economists that financial incentives affect performance for the better (e.g., Davis & Holt 1993; Harrison 1992; Roth 1995; Smith 1991; Smith & Walker 1993a; 1993b). Consequently, experimental economists have debated the "growing body of evidence [from psychology] – mainly of an experimental nature – that has documented systematic departures from the dictates of rational economic behavior" (Hogarth & Reder 1987, p. vii; see e.g., Kahneman et al. 1982; Kahneman & Tversky 1996; Tversky & Kahneman 1981), often on the grounds that such departures have been shown primarily in experiments without financial incentives (e.g., Smith 1991, p. 887).

The rationale behind this criticism is that economists think of "cognitive effort" as a scarce resource that people have to allocate strategically. If participants are not paid contingent on their performance, economists argue, then they will not invest cognitive effort to avoid making judgment errors, whereas if payoffs are provided that satisfy saliency and dominance requirements (Smith 1976; 1982; see also Harrison 1989 and 1992[5]), then "subject decisions will move closer to the theorist's optimum and result in a reduction in the variance of decision error" (Smith & Walker 1993a, p. 260; there is an interesting link to the psychology studies on the relationship between "need for cognition" and the quality of decision making: see, e.g., Smith & Levin 1996). Believers in the reality of violations of rational economic behavior in both psychology and economics have dismissed this criticism (e.g., Thaler 1987; Tversky & Kahneman 1987).

Our 10-year sample of empirical studies published in *JBDM* was not selected to demonstrate whether financial incentives matter; therefore it can add systematic empirical evidence. Recall that in our sample of *JBDM* studies, 48 of 186 studies (26%) employed financial incentives. In only 10 of those 48 studies, however, was the effect of payments systematically explored, either by comparing a payment to a nonpayment condition or by comparing different payment schemes. What results were obtained in those 10 studies?

For the studies in which the necessary information was given, we calculated the effect size *eta,* which can be defined as the square root of the proportion of variance accounted for (Rosenthal & Rosnow 1991). *Eta* is identical to the Pearson product-moment correlation coefficient when $df = 1$, as in the case when two conditions are being compared. According to Cohen's (1988) classification of effect sizes, values of *eta* of .1, .3, and .5 constitute a small, medium, and large effect size, respectively. As can be seen in Table 2, the effect sizes for financial incentives ranged from small to (very) large, confirming findings in other review studies (e.g., Camerer & Hogarth 1999).

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Table 2. *Effects of performance-based payments obtained in studies from* Journal of Behavioral Decision Making. *We calculated effect size* eta *and* d *(Cohen 1988; Rosenthal & Rosnow 1991) when sufficient information was available*

Authors	Task	Comparison	General results (effect size)
Levin, Chapman & Johnson (1988)	Judgments of the likelihood of taking gambles, and expressing confidence in those judgments. Probability information was framed either positively as "percent chance of winning" or negatively as "percent chance of losing"	*Study 1* Hypothetical gambles *Study 2* Gambles with real money; winnings ranged from 25¢ to $5, averaging about $2	Reduced framing effect, that is, difference in likelihood of taking the gamble depending on positive vs. negative framing was smaller in real-money gambles than in hypothetical gambles;[a] in both studies, confidence ratings were higher for incomplete than for complete information trials in the explicit inference condition.
Allison & Messick (1990)	Division of shared resource: Participants decide how much they request from a resource, which they share with others	*Low-payoff condition* Resource $9, $10.50 *High-payoff condition* Resource $12, $13.50	No "significant" main effect of payoff condition (p. 201); however, the proportion of the pool requested was on average 24.1% and 29.1% (see Allison & Messick's Table 1) in the low- and high-payoff condition, respectively. Moreover, proportion of requested resource was closest to normative model in the high-payoff condition, in which the pool was nondivisible and other group members could not punish self-regarding behavior (45.8%).[b]
Irwin, McClelland & Schulze (1992)	Gambles in which participants face a 1% chance of a $40 loss. In an auction procedure, they could enter (optimal) bids for insurance against this loss	*Hypothetical reward* *Real reward* Participants received an initial stake of $50 or $80 and kept the positive balance at the conclusion	For hypothetical reward, the bids (in terms of σ) were more variable (*eta* = .60). There were more very high bids (*eta* = .46) and more "0" bids (*eta* = .62) for the insurance in the hypothetical- than in the real-reward condition. In addition, bids were more optimal (measured in terms of the difference between expected value of the insurance and bids) in the real-reward than in the hypothetical-reward condition (difference: *eta* =.39; absolute difference: *eta* = .45).
Van Wallendael & Guignard (1992)	Categorization of stimuli with the aid of information of varying costs and diagnosticity levels[c]	*Study 1* Hypothetical points for correct identification, and hypothetical cost for each piece of information *Study 2* $1.25 for every correct identification, minus 5¢, 10¢, and 15¢ per piece of information	A main effect of information diagnosticity (inconsistent with the expected-value approach) was larger in Study 2 (*eta* = .79) than in Study 1 (*eta* = .21); no "significant" main effect for information costs in Study 1 but a "significant" main effect consistent with the expected-value approach in Study 2 (*eta* = .70). In both studies, participants expressed greater confidence in decisions based on high-diagnosticity questions than on low-diagnosticity questions; this effect, however, was larger in Study 2 than in Study 1 (*eta* = .24, *eta* = .74). In both studies, there were no significant differences in confidence owing to answer diagnosticity or information costs, nor significant interactions.
Hulland & Kleinmuntz (1994)	Choice of preferred alternative based upon searched information	*Incentive condition* Incentive to spend more effort (one could win a year's subscription to a chosen alternative among women's magazines) *Control condition* Nothing at stake	According to an analysis of variance, the effects of incentive on various search measures were not significant (p. 87). In a "partial least square" analysis, the authors found that in the incentive condition "more external search and internal (non-search) processing and . . . significantly more effort overall" (p. 97) was expended.[d]
Van Wallendael (1995)	Judgments of the guilt or innocence of suspects in fictional crimes (based on the purchase of infor-	*Study 2* Hypothetical costs for information and wrong decisions,	"The monetary prize had virtually no effect on subjects' performance. Results of Experiment 3 replicated all of the major findings of Experiment 2" (p. 259). For instance, participants in

(continued)

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Table 2. *(Continued)*

Authors	Task	Comparison	General results (effect size)
	mation) and probability ratings of those judgments being correct	and hypothetical payoffs for correct decisions *Study 3* Real costs for information and real payoff for correct decisions (the highest scorer among 10 scorers wins $10)	both studies overbought in comparison to optimal buys calculated from an expected-value model.[e]
Mellers, Berretty & Birnbaum (1995)	Evaluation of the worth of gambles[f]	*Study 2* Hypothetical prices in the *baseline* condition; in the *incentives* condition a pair of gambles was randomly selected, and the gamble that was preferred was actually played	"There are no significant main effects or interactions involving financial incentives" (p. 210). However, the mean percentage of violations, nonviolations, and ties across 10 gamble pairs was strongly affected. In the baseline condition 38%, 19%, and 43% were violations, nonviolations, and ties, whereas the proportions in the incentives condition were 36%, 50%, and 16% (compare Exhibits 7 and 8, similar results for Exhibits 11 and 12). If one only considers violations and nonviolations, then in all 10 gamble pairs in Exhibit 7 (baseline condition), the proportion of violations exceeded that of nonviolations. In contrast, in Exhibit 8 (incentives condition) this is only true for 2 of the 10 gamble pairs.
Ordóñez, Mellers, Chang & Roberts (1995)	Evaluation (attractiveness ratings, strength of preference judgments, selling prices) of pairs of gambles in a sequential vs. simultaneous way (i.e., evaluate pair of gambles on two types of response modes before evaluating the next pair)	*No financial incentives* *Financial incentives* A pair of gambles was randomly selected, and the gamble that was preferred was actually played (payments ranged from $0 to $11.34)	Without incentives, only in one of the three combinations of response modes were reversals "significantly" reduced (*eta* = .27) when comparing sequential to simultaneous performance. With incentives, reversal rates were "significantly" reduced in all three combinations of two response modes (*eta* = .33, .49, .51) when comparing sequential to simultaneous performance. Ordóñez et al. concluded that "simultaneous performance with financial incentives virtually eliminated reversal rates" (p. 271–72), and "although preference reversals are quite robust, the present results show that subjects can give consistent responses with certain procedures and incentives" (p. 276).
Yaniv & Schul (1997)	Selection of answers from a set of multiple alternatives under two different framings: *Include* likely alternatives from initial set versus *exclude* the least likely alternatives	*Study 1* No payment *Study 2* Payoff rule (which reflected a tradeoff between the number of marked alternatives and probability that the correct choice was marked)	The accuracy of participants' selections is almost identical in Studies 1 and 2 (50% vs. 49%). Although exclusion and inclusion conditions are still significantly different, the difference is reduced in Study 2: The difference in the size of the choice set is larger in Study 1 than in Study 2. Expressed in percentage of the full set, the difference between the inclusion and the exclusion set is 31.5 (d = 3.2) percentage points in Study 1 and 20 percentage points (d = 1.9) in Study 2. In addition, the difference in accuracy between the inclusion and the exclusion condition is smaller in Study 2 (18 percentage points, d = 1.2) than in Study 1 (38 percentage points, d = 2.7).

(continued)

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Table 2. *(Continued)*

Authors	Task	Comparison	General results (effect size)
Beeler & Hunton (1997)	Allocation decisions into investment portfolios, and the effect of negative feedback on performance of commitment and information search	*No-pay condition and Salary condition* Choice of two state lottery tickets or $2, irrespective of how well their investments performed *Contingent compensation* Earning of state lottery tickets or cash based on investment performance	As a function of compensation methods, an escalation of commitment ("sunk cost effect") was observed, that is, the average dollars subsequently invested in losing companies was $35,890 (no-pay), 64,280 (salary), and $119,450 (contingent). Amount of time participants viewed prospective information decreased and amount of time for retrospective information increased as a function of compensation method (prospective time: 360.0, 258.5, and 151.5 seconds; retrospective time: 218.5, 346.0, and 469.0 seconds).[g]

[a]Levin et al. concluded that "high levels of personal involvement, such as created by providing real monetary consequences to gambles, can serve to . . . reduce the information framing effect" (p. 39).

[b]As one normative model, Allison and Messick (p. 197) assumed that each participant should have requested all that was permissible, leaving a minimal amount for the final group member. Based on their results, they concluded that "subjects in social decision tasks involving shared resources cannot be modeled as strategic money maximizers" (p. 195). We point out, however, that there is a major methodological problem in this study, namely that of the lack of control of the "social distance" between experimenter and participant. Confronted with persistent deviations from game theoretic predictions in dictator and ultimatum games, Hoffman et al. (1996) manipulated instructional and procedural aspects of the design and implementation of dictator games and found that increasing social distance – in the most extreme case, no one, including the experimenter or any observer of the data, could possibly know any participant's decision (i.e., complete anonymity) – monotonically reduces the deviations from game-theoretic predictions. In Allison and Messick's study, anonymity was guaranteed neither between participant and experiment nor among participants ("each member would be told what the previous members took," p. 199).

[c]As a normative model, Van Wallendael and Guignard assumed that if participants used an expected-value approach their information purchases should show main effects of cost (with more information being purchased at lower costs) and question diagnosticity (with more high-diagnosticity information being purchased), and other interaction effects. However, they also noted that an alternative model, which better predicted people's information purchases is "not necessarily normatively suboptimal" (p. 36).

[d]Hulland and Kleinmuntz pointed out that the link between decisions and their consequences may have been small, as many male participants may have been unaffected by the incentive manipulation (subscription to a women's magazine). They suggest that this may explain why incentives did not affect another search measure (participants' use of summary evaluations).

[e]We will not report effect sizes for the multiple statistical comparisons (the study included a $2 \times 2 \times 3$ factorial design and two dependent measures).

[f]As a normative model, Mellers et al. assumed the dominance principle, which holds that increasing one or more outcomes of a gamble should increase the judged price of the gamble, with everything else held constant.

[g]The authors concluded from their results that "performance-based incentives, (i.e., contingent) led to higher escalation of commitment," and that "it appears as though individuals who heightened their commitment to an initial course of action endeavored to resolve cognitive dissonance, justify past decisions, and account for prior actions by searching for retrospective, supporting information" (p. 88). Note, however, that all participats received the *false* information that the data they studied prior to their investment decisions represented actual cases, and they also were given the feedback, independent of their specific investment decision, that their investment portfolio had declined in value. If participants in the contingent-compensation condition spent more cognitive effort prior to their decision, their search of retrospective investment may not express their increasing commitment but their potential disbelief concerning the feedback fabricated by the experimenter.

In the majority of cases where payments made a difference, they improved people's performance. Specifically, payments decreased a framing effect (Levin et al. 1988), made people take the cost of information into account, and increased their confidence in decisions based on highly diagnostic information (Van Wallendael & Guignard 1992). In an auction experiment, payments brought bids closer to optimality and reduced data variability (Irwin et al. 1992). Payments also decreased the percentage of ties in gamble evaluations relative to nonviolations of the dominance principle (Mellers et al. 1995) and, when combined with "simultaneous" judgment, eliminated preference reversals (Ordóñez et al. 1995). In addition, payments reduced the noncomplementarity of judgments (Yaniv & Schul 1997), brought people's allocation decisions closer to the prescriptions of an optimal model (when self-regarding behavior could be punished; Allison & Messick 1990), and induced people to expend more effort (in terms of external search and internal nonsearch processing) in making choices (Hulland & Kleinmuntz 1994). In only two cases did payments seem to impair performance: They escalated commitment and time spent obtaining retrospective information (sunk cost effect, Beeler & Hunton 1997; but see the methodological problems mentioned in Table 2, footnote g) and ac-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

centuated a (suboptimal) information diagnosticity effect (Van Wallendael & Guignard 1992).

In a few cases, payments did not make a difference. As Table 2 shows, they did not improve either confidence judgments (Levin et al. 1988; Van Wallendael & Guignard 1992) or patterns of information purchase and probability ratings based on that information (Van Wallendael 1995). They also did not decrease the proportion of violations of the dominance principle (Mellers et al. 1995), nor did they increase the accuracy of participants' responses to general knowledge items (Yaniv & Schul 1997).

Given that Table 2 reports *all* studies of the *JBDM* sample that systematically explored the effect of financial incentives, we conclude that, although payments do not guarantee optimal decisions, in many cases they bring decisions closer to the predictions of the normative models. Moreover, and equally important, they can reduce data variability substantially. These results are in line with Smith and Walker's (1993a) survey of 31 experimental studies reporting on the effects of financial incentives and decision costs (including, e.g., Grether & Plott's 1979 study of preference reversals). Specifically, Smith and Walker (1993a) concluded that "in virtually all cases rewards reduce the variance of the data around the predicted outcome" (p. 245, see further evidence in Grether 1980; Harless & Camerer 1994; Jamal & Sunder 1991).

Aside from the Smith and Walker study, four other recent review articles have explored the effect of financial incentives. First, Camerer and Hogarth (1999) reviewed 74 studies (e.g., on judgment and decision making, games, and market experiments) and compared the behavior of experimental participants who did and did not receive payments according to their performance. Camerer and Hogarth found cases in which financial incentives helped, hurt, did not make a difference, and made a difference although it was not clear whether for better or worse because there was no standard for optimal performance. More specifically, however, Camerer and Hogarth found that financial incentives have the largest effect in "judgment and decision" studies – our focus and running example of the sharply differing practices between experimental economists and psychologists: Out of 28 studies, in 15, financial incentives helped, in 5, they did not have an effect, and in 8, they had negative effects. Regarding the latter, however, Camerer and Hogarth wrote that the "effects are often unclear for various methodological reasons" (p. 21). Moreover, Camerer and Hogarth reported that in many of those studies in which incentives did not affect mean performance, they "*did* reduce variation" (p. 23, their emphasis).

Second, Harrison and Rutstroem (in press), drawing on 40 studies, accumulated overwhelming evidence of a "hypothetical bias" in value elicitation methods. Simply put, they found that when people are asked hypothetically what they would be willing to pay to maintain an environmental good (e.g., the vista of the Grand Canyon), they systematically overstate their true willingness-to-pay (see also Harrison, 1999, for a blunt assessment and methodological discussion of the state of the art of contingent valuation studies). Camerer and Hogarth (1999) mentioned the Harrison and Rutstroem study briefly under the heading "When incentives affect behavior, but there is no performance standard." We believe this to be a misclassification. In our view, true willingness-to-pay is a norm against which "cheap talk" can be measured.

Third, in a meta-analytic review of empirical research (from several applied psychology journals), Jenkins et al. (1998) found financial incentives to be related to performance quantity (e.g., exam completion time) but not quality (e.g., coding accuracy; the authors stressed that this result ought to be "viewed with caution because it is based on only six studies," p. 783). They found an effect size for performance quantity of .34 (point-biserial correlation), which is considered to be of medium size (e.g., Rosenthal & Rosnow 1991). In addition, they reported that the relation between financial incentives and performance is weakest in laboratory experiments (as compared, e.g., to field experiments) – possibly because "laboratory studies typically use small incentives" (Jenkins et al. 1998, p. 784). While their review does not address the impact of financial incentives on intrinsic motivation directly, they concluded that "our results . . . go a long way toward dispelling the myth that financial incentives erode intrinsic motivation" (p. 784). Fourth, and of relevance in light of Jenkins et al.'s results, Prendergast (1999) reviewed the effect of incentive provision in firms and found that there is a positive relationship between financial incentives and performance.

To conclude, concerning the controversial issue of the effects of financial incentives, there seems to be agreement on at least the following points: First, financial incentives matter more in some areas than in others (e.g., see Camerer & Hogarth's 1999, distinction between judgment and decision vs. games and markets). Second, they matter more often than not in those areas that we explore here (in particular, research on judgment and decision making), which are relevant for both psychologists and economists. Third, the obtained effects seemed to be two-fold, namely, convergence of the data toward the performance criterion and reduction of the data's variance. Based on these results, we propose that psychologists in behavioral decision making consider using financial incentives. Although "asking purely hypothetical questions is inexpensive, fast and convenient" (Thaler 1987 p. 120), we conjecture that the benefits of being able to run many studies do not outweigh the costs of generating results of questionable reliability (see also Beattie & Loomes 1997, p. 166).

In addition, only by paying serious attention to financial incentives can psychologists conduct systematic research on many open issues. (For instance, under which conditions do financial incentives improve, not matter to, or impair task performance? For previous research on these conditions, see, e.g., Beattie & Loomes 1997; Hogarth et al. 1991; Payne et al. 1992; Pelham & Neter 1995; Schwartz 1982; Wilcox 1993.)[6] How do incentives (and opportunity costs) affect decision strategies and information processing (e.g., Payne et al. 1996; Stone & Schkade 1994; Wallsten & Barton 1982;), and how do they interact with other kinds of incentives (e.g., social incentives) and motives?[7] Some of the reported research also highlights the need to understand better how incentives interact with other variables of experimental design (e.g., repetition of trials, Chu & Chu 1990, and presentation of gambles, Ordóñez et al. 1995; see also Camerer 1995, sect. I and Camerer & Hogarth 1999), and to establish what kinds of salient and dominant rewards are effective (e.g., the problem of flat maxima, see Harrison 1994; von Winterfeldt & Edwards 1982).

Ultimately, the debate over financial incentives is also an expression of the precision of theories or a lack thereof. Economists virtually always pay because the explicit do-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

main of economic theories is extrinsically motivated economic behavior. Psychological theories in behavioral decision making often do not make it completely clear what behavior they target – intrinsically or extrinsically motivated behavior. If theories were more explicit about their domain and the implicated motivation, then they rather than the experimenters would define the appropriate test conditions.

We conclude this section by briefly discussing two possible reasons for mixed results (sect. 4.1.1), and whether and how payments affect intrinsic motivation (sect. 4.1.2).

4.1.1. Reasons for the mixed results? The majority of the results in Table 2 are inconsistent with studies that did not find any effect of payment (see, e.g., the studies mentioned in Dawes 1988; Hogarth et al. 1991; Stone & Ziebart 1995). How can these discrepant results be explained? There are at least two *possible* explanations. The first was pointed out by Harrison (1994, p. 240), who reexamined some of Kahneman and Tversky's studies on cognitive illusions that used financial incentives and concluded that the majority of these experiments lack payoff dominance (see note 5). In other words, not choosing the theoretically optimal alternative costs participants in these experiments too little. Based on new experiments (e.g., on preference reversals and base-rate neglect) that were designed to satisfy the dominance requirement, Harrison (1994) concluded that in his redesigned experiments, observed choice behavior is consistent with the predictions of economic theory.

A second possible explanation can be drawn from the existence of multiple and contradictory norms against which performance might be compared (see, e.g., the controversy between Kahneman & Tversky 1996 and Gigerenzer 1996; see also Hilton 1995 on the issue of conversational logic). The problem of multiple and ambiguous norms may be compounded by a focus on coherence criteria (e.g., logical consistency, rules of probability) over correspondence criteria, which relate human performance to success in the real world (e.g., speed, accuracy, frugality). Clearly, if multiple norms exist and the experimenter does not clarify the criterion for which participants should aim (e.g., by specification of payoffs), then payment will not necessarily bring their responses closer to the normative criterion the experimenter has in mind. More generally, as argued by Edwards (1961):

> Experiments should be designed so that each subject has enough information to resolve ambiguities about how to evaluate the consequences of his own behavior which are inherent in conflicting value dimensions. That means that the subject should have the information about costs and payoffs . . . necessary to evaluate each course of action relative to all others available to him. (p. 283)

4.1.2. How do financial incentives affect intrinsic motivation? An important argument against the use of financial incentives is that they crowd out intrinsic motivation (if it exists). This argument can be traced back to Lepper et al.'s (1973) finding that after being paid to perform an activity they seemed to enjoy, participants invested less effort in the activity when payoffs ceased. Lepper et al. interpreted participants' initial apparent enjoyment of the activity as evidence of intrinsic motivation and their subsequent decrease in effort expenditure as evidence of the negative impact of extrinsic rewards on intrinsic motivation. A huge literature has evolved consequently. Drawing on an extensive meta-analysis by Cameron and Pierce (1994), Eisenberger and Cameron (1996) performed a meta-analysis on the question of whether financial incentives really undermine intrinsic motivation.

Based on their examination of two main measures of intrinsic motivation, namely, the free time spent on the task post-reward and the expressed attitude toward the task, they did not find that completion-dependent reward (i.e., reward for completing a task or solving a problem) had any negative effect. Moreover, they found that quality-dependent reward (i.e., reward for the quality of one's performance relative to some normative standard) had a positive effect on expressed attitudes toward the task. Ironically, the only measure on which Eisenberger and Cameron (1996) found a reliable negative effect was the free time spent carrying out the activity following performance-*independent* reward (i.e., reward for simply taking part in an activity), the type of reward commonly used in psychological experiments. Eisenberger and Cameron (1996) concluded that "claimed negative effects of reward on task interest and creativity have attained the status of myth, taken for granted despite considerable evidence that the conditions producing these effects are limited and easily remedied" (p. 1154).

The conclusions of Cameron and colleagues have been challenged (e.g., Deci et al. 1999a; Kohn 1996; Lepper et al. 1996; see also the debate in the *American Psychologist,* June 1998). In their most recent meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation, Deci et al. also discussed the procedure employed by Eisenberger and Cameron (1996). No surprise, they come to very different conclusions, confirming the classic finding that tangible rewards (i.e., financial incentives) undermine intrinsic motivation. One important bone of contention is the definition of the relevant set of studies. Deci et al. argue that it ought to be confined to "interesting" tasks, and ought to exclude "boring" tasks, some of which Eisenberger and Cameron include (see also Deci et al. 1999b; Eisenberger et al. 1999; Lepper et al. 1999). In sum, there is agreement that rewards can be used as a technique of control; disagreement exists as to unintended consequences of rewards. We believe that the situation calls for a meta-analysis done by the two camps and a jointly determined arbiter following the model of "adversarial collaboration" proposed by Kahneman and exemplified in Mellers et al. (2001). In the meantime, we believe that the boring nature of many experiments and the available evidence reported here suggest that financial incentives matter in tasks examined in behavioral decision making (see Table 1; Camerer & Hogarth 1999) and thus ought to be considered, unless previous studies show that financial incentives do not matter for a particular task.[8]

5. Honesty versus deception

Deceiving participants is generally taboo among experimental economists (Davis & Holt 1993, p. 24) and, indeed, economics studies that use deception can probably be counted on two hands.[9] Davis and Holt (1993, pp. 23–24; see also, Hey 1991; Ledyard 1995) gave the following typical rationale for economists' reasons to argue against deception (for a rare dissenting view in economics, see Bonetti 1998, but see also the comments of Hey 1998; McDaniel & Starmer 1998):

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

> The researcher should . . . be careful to avoid deceiving participants. Most economists are very concerned about developing and maintaining a reputation among the student population for honesty in order to ensure that subject actions are motivated by the induced monetary rewards rather than by psychological reactions to suspected manipulation. Subjects may suspect deception if it is present. Moreover, even if subjects fail to detect deception within a session, it may jeopardize future experiments if the subjects ever find out that they were deceived and report this information to their friends.

Even if participants initially were to take part in experiments out of a sense of cooperation, intrinsic motivation, or the like, economists reason that they will probably become distrustful and start second-guessing the purpose of experiments as soon as they hear about such deception. In other words, economists fear reputational spillover effects of deceptive practices even if only a few of their tribe practice it. In the parlance of economists, participants' expectation that they will not be deceived (i.e., honesty on the part of the experimenter) is a common good of sorts (such as air or water) that would be depleted (contaminated) quickly if deception was allowed and the decision about its use left to each experimenter's own cost-benefit analysis. On theoretical and empirical grounds, economists do not trust experimenters to make an unbiased analysis of the (private) benefits of deception and its (public) costs. The temptation, or, in economists' parlance, the "moral hazard" to capture the private benefits of deception is perceived to be simply too strong. Indeed, given that the American Psychological Association (APA) ethics guidelines (APA 1992, p. 1609) propose to employ deception as a last-resort strategy, to be used only after careful weighing of benefits and costs, the frequent use of deception in some areas of psychology seems to confirm economists' fear.

Take the highest ranked journal in social psychology, the *Journal of Personality and Social Psychology* (*JPSP*), and its predecessor, *Journal of Abnormal and Social Psychology,* as an illustration. After a sharp upswing during the 1960s (where it tripled from 16% in 1961 to 47% in 1971), the use of deception continued to increase through the 1970s, reaching its high in 1979 (59%) before dropping to 50% in 1983 (Adair et al. 1985). Since then it has fluctuated between 31% and 47% (1986: 32%, 1992: 47%, 1994: 31%, 1996: 42%; as reported in Epley & Huff 1998; Nicks et al. 1997; Sieber et al. 1995).

While some of these fluctuations may reflect different definitions of what constitutes deception (e.g., compare the more inclusive criteria employed by Sieber et al. with the criteria used by Nicks et al.), a conservative estimate would be that every third study published in *JPSP* in the 1990s employed deception. (In other social psychological journals, e.g., *Journal of Experimental Social Psychology,* the proportion is even higher; Adair et al. 1985; Nicks et al. 1997.) The widespread use of deception in social psychology in recent years contrasts markedly with its decidedly more selective use in the 1950s and earlier (Adair et al. 1985). Although deception is likely to be most frequent in social psychology, it is not restricted to it (see sects. 6.1 and 6.3 in the discussion).

Why do psychologists use deception? Although some critics of the frequent use of deception attributed it to a "fun-and-games approach" (Ring 1967, p. 117) to psychological experimentation, today's primary motivation for deception seems to rest on at least two serious methodological arguments: First, if participants were aware of the true purpose of a study, they might respond strategically and the investigator might lose experimental control. For instance, one might expect participants to "bend over backwards" (Kimmel 1996, p. 68) to show how accepting they are of members of other races if they know that they are participating in a study of racial prejudices. To the extent that psychologists, more than economists, are interested in social behavior and "sensitive" issues, in which knowledge of the true purpose of a study could affect participants' behavior (e.g., attitudes and opinions), one might expect deception to be used more often in psychology. The second argument is that deception can be used to produce situations of special interest that are unlikely to arise naturally (e.g., an emergency situation in which bystander effects can be studied).

Despite "widespread agreement" that deception is "often a methodological necessity" (Kimmel 1996, p. 68), and the claim that there is no reason to worry about the methodological consequences of deception (e.g., Christensen 1988; Sharpe et al. 1992; Smith & Richardson 1983), its use has been a longstanding and persistent concern in psychology. Anticipating economists' common good argument, Wallsten (1982) suggested that the erosion of participants' trust would hurt everyone who relies on the participant pool. While some authors proposed cosmetic changes in the use of deception (e.g., Taylor & Shepperd 1996), others proposed more drastic measures (e.g., Baumrind 1985; Kelman 1967; MacCoun & Kerr 1987; Newberry 1973; Ortmann & Hertwig 1997; 1998; Schultz 1969; Vinacke 1954).

5.1. Does deception matter?

Our concern here is pragmatic not ethical (see Baumrind 1964; 1971; 1985), that is, we are interested in the methodological consequences of the use of deception on participants' attitudes, expectations, and in particular, on participants' behavior in experiments. Before we discuss the available evidence, it is useful to conceptualize the interaction between participant and experimenter as a one-sided prisoner's dilemma, or principal-agent game. Such a game models the relationship between an agent and a principal, both of whom can either contribute their respective assets (trust for the principal, honesty for the agent) or withhold them. In the current context, the experimenter (agent) can choose either to deceive participants or to be truthful about the setting and purpose of the experiment, while the participant (principal) can choose either to trust the experimenter or to doubt the experimenter's claims. The game-theoretic predictions for a one-shot principal-agent game are, dependent on the parameterization, clear-cut: The agent will defect – at least with some probability. The principal, anticipating the defection, will doubt the experimenter's claims – at least with some probability (see Ortmann & Colander 1997, for two typical parameterizations).

The interaction between agent and principal, of course, is not likely to be a one-shot game. Participants (principals) may come into the laboratory either inexperienced or experienced (by way of previous participation in deception experiments). If they are experienced, then that experience may bear directly on their expectation of the experimenter's action choice. If they are inexperienced, then other participants' experience may still bear on their expectation. If participants have reason to trust the experimenter, they may

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

act like the "good" (Orne 1962) or "obedient" (Fillenbaum 1966) participants they are often assumed to be in psychology (see Rosenthal & Rosnow 1991). If they have reason to believe that the agent will deceive them, however, their behavior may range from suspicious to apathetic (Newberry 1973) and negativistic (Christensen 1977; Weber & Cook 1972).

Experimental results from trust games suggest that people (participants) may accept being fooled once, but not twice (Dickhaut et al. 1995). Recent results reported by Krupat and Garonzik (1994) also suggest that prior experience with deception affects participants' expectations, that is, increases their suspicion (see also Epley & Huff 1998). According to Krupat and Garonzik (1994), such suspicion is likely to introduce "considerable random noise" into their responses (p. 219). In this context it is interesting to note that Stang (1976) already pointed out that the percentage of suspicious participants (in conformity experiments) tracked closely the increase of the use of deception through the 1960s.

Ironically, the APA ethical guidelines concerning debriefing may exacerbate rather than diminish participants' suspicion: "Deception that is an integral feature of the design and conduct of an experiment must be explained to participants as early as it is feasible, preferably at the conclusion of their participation, but no later than at the conclusion of the research" (APA 1992, p. 1609). From an ethical point of view, debriefing is the right thing to do; from a pragmatic point of view, however, it only undermines the trust of actual and potential participants and thereby contaminates the data collected in future experiments: "Each time this quite proper moral requirement is met, the general impression that psychologists commonly deceive is strengthened" (Mixon 1972, p. 145).

Notwithstanding this concern regarding the use of deception, a number of researchers in psychology have advocated its use on the grounds that participants have a favorable attitude toward it. Smith and Richardson (1983), for example, observed that participants in experiments involving deception reported having enjoyed, and indeed having benefited from, the experience more than those in experiments without deception. Summing up his review of research on the impact of deception on participants, Christensen (1988) concluded: "This review . . . has consistently revealed that research participants do not perceive that they are harmed and do not seem to mind being misled. In fact, evidence exists suggesting that deception experiments are more enjoyable and beneficial than nondeception experiments" (p. 668). In Christensen's (1988) view, "the scale seems to be tilted in favor of continuing the use of deception in psychological research" (p. 664; see also Aitkenhead & Dordoy 1985; Sharpe et al. 1992).

However, even if undergraduate participants tell experimenters (often their professors) the truth about how they feel about deception and genuinely do not mind it (Smith & Richardson 1983), which is by no means a universal finding (e.g., Allen 1983; Cook et al. 1970; Epstein et al. 1973; Fisher & Fyrberg 1994; Rubin 1985; Oliansky 1991), we believe that studies of feelings about and attitudes toward deception overlook a key issue, namely, the extent to which deception affects participants' *behavior* in experiments. Some intriguing findings suggest that, ironically, it is sometimes the experimenter who is duped in an experiment employing deception. For example, Newberry (1973) found that a high percentage of participants, given a tip-off by an experimental confederate, do not admit to having had foreknowledge when questioned later (30–80% in various conditions) – a result that surely undermines the frequent assumption that participants are cooperative (e.g., Bröder 1998; Kimmel 1998).

MacCoun and Kerr (1987) gave a particularly dramatic example that indicates that participants' behavior is affected by the expectation of deception: When a participant had an epileptic seizure during an experiment, the other participants present appeared to believe the seizure was a charade perpetrated by the experimenter and a confederate and therefore initially ignored it. The only person who immediately helped the victim was the only one who had no prior psychology coursework (MacCoun & Kerr 1987). Along the same lines, Taylor and Shepperd (1996) conducted an experiment in which they used deception to study the effectiveness of conventional debriefing procedures in detecting suspicion of deception. Despite explicit instruction not to communicate while the experimenter left the room on a pretext, participants talked during the experimenter's absence and thereby found out that they were being deceived. In a debriefing, none of them revealed this discovery.

To conclude, because psychology students are the main data source in psychological studies (Sieber & Saks 1989), a substantial proportion of participants can be expected to have experienced deception directly. Owing to students' general expectations (due to coursework) or direct personal experiences, deception can have (negative) consequences even in those domains of psychology in which deception is not or is less frequently used. We therefore concur with the argument advanced by economists and (some) psychologists that participants' trust is a public good worth investing in to increase experimental control. We propose that psychologists view the use of deception as involving a trade-off not only "between methodological and ethical considerations" (Kimmel 1996, p. 71), but also between its methodological costs and benefits.

6. General discussion

In this article, we have been concerned with practices of psychological experimentation and their divergence from those of experimental economics. In particular, we considered four key variables of experimental design that take on markedly different realizations in the two disciplines. We argued that the conventions in economics of providing and having participants enact a script, repeating trials, giving financial incentives, and not deceiving participants are *de facto* regulatory, allowing for comparatively little variation in experimental practices between researchers. The corresponding experimental practices in psychology, by contrast, are not regulated by strong conventions. This laissez-faire approach allows for a wide range of experimental practices, which in turn may increase variability in the data obtained and ultimately may impede theoretical advances.

Are our findings consonant with psychologists' and economists' perceptions of their own and the other discipline's practices? Why do we see different realizations of key variables across different disciplines and what are the policy

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

implications of our arguments? In the next sections, we address each of these questions in turn.

6.1. How researchers describe their own practices and those of the other discipline

We have provided various illustrations for the two theses we proposed, namely, that (1) key variables of experimental design tend to be realized differently in economics and psychology and (2) experimental standards in economics are regulatory in that they allow for little variation between the experimental practices of individual researchers, whereas experimental standards in psychology are comparatively laissez-faire.

Are these two theses also reflected in the way experimentalists in both fields describe their own practices? We conducted a small-scale survey in which we asked researchers in the fields of behavioral decision making and experimental economics to respond to nine questions concerning the use of financial incentives, trial-by-trial feedback, and deception. The questions asked researchers to describe their own research practices (e.g., "How often do you use performance-contingent payments in your experiments?"), research practices in their field generally (e.g., "How often do you think that experimenters in economics/JDM research use performance-contingent payments?"), and research practice in the related field (e.g., "How often do you think that experimental economists/psychologists use performance-contingent payment?"). Researchers were asked to provide their responses in terms of absolute frequencies ("In __ out of 10 experiments?"); alternatively, they could mark an "I don't know" option.

We sent the questionnaire to the electronic mailing lists of the *European Association for Decision Making* and the *Brunswik Society.* Both societies encompass mostly European and American psychologists interested in judgment and decision making. We also distributed the questionnaire at the 1999 annual meeting of the *Economic Science Association,* which is attended by experimental economists. A total of 26 researchers in psychology and 40 researchers in economics responded. Admittedly, the response rate for psychologists was quite low (the response rate for economists was about 60%); both samples, however, encompassed well-established as well as young researchers.

Economists estimated that, on average, they used financial incentives in 9.7 out of 10 experiments ($MD = 10$, $SD = .8$); trial-by-trial feedback in 8.7 out of 10 experiments ($MD = 9$, $SD = 2.1$), and deception in .17 out of 10 experiments ($MD = 0$, $SD = .44$). In contrast, psychologists' average estimates were 2.9 for financial incentives ($MD = 1$, $SD = 3.5$), 2.4 for trial-by-trial feedback ($MD = 1$, $SD = 3.2$), and 1.7 for deception ($MD = 0$, $SD = 2.8$). Aside from the drastically different self-reported practices across fields, the results also demonstrate the wider range of practices within psychology. Concerning financial incentives, for instance, 40% of psychologists responded that they never use financial incentives, whereas 32% use them in half or more of their experiments. Regarding deception, 60% stated that they never use it, whereas 20% use it in half or more of their experiments. When we asked researchers to characterize the general practices in their own field on the same measures, we obtained responses close to those described above. However, researchers in both groups believed that they use financial incentives and trial-by-trial feedback slightly more often and deception slightly less often than researchers in their field as a whole.

To what extent are psychologists and economists aware that experimental practices are different in the other field? Although the psychologists were aware that practices in economics differ from those in their own field, they underestimated the extent of the differences. On average, they estimated that economists use financial incentives in 5.6 out of 10 experiments, give trial-by-trial feedback in 3.2 out of 10 experiments, and use deception in 1.2 out of 10 experiments. Although economists' estimates of the use of financial incentives by psychologists was fairly accurately calibrated ($M = 2.3$), they overestimated the use of trial-by-trial feedback ($M = 4.5$) and deception ($M = 5.5$) by psychologists.[10]

The results of our small-scale survey are consistent with the two theses we proposed: Experimental practices in behavioral decision making and economics differ and the research practices of psychologists are much more variable. Although some of this variability is likely to be driven by behavioral decision making researchers' interest in questions that do not lend themselves to the use of financial incentives or trial-by-trial feedback, we suggest that the large variance in their responses also reflects the lack of standards committing them to consistency in experimental practices.

6.2. Why do the methodological practices differ?

There is no simple answer to this question. Differences in experimental practices are neither recent nor confined to cross-disciplinary comparisons. Danziger (1990) identified at least three diverging models of investigative practice in early modern psychology: the Wundtian, the clinical, and the Galtonian. According to Danziger, the investigators' different research goals drove different practices. Whether one wanted to learn about pathological states (French investigators of hypnosis), individual differences (Galton), or elementary processes in the generalized human mind (Wundt) determined what investigative situations seemed appropriate. Researchers in contemporary psychology pursue a multitude of research goals as well, and not only those of early modern psychology. To the extent that Danziger's (1990) thesis that different goals give rise to different investigative practices is valid, the heterogeneity of experimental practices within psychology therefore should not be surprising.[11]

In contrast to psychology, experimental economics displays much less variability in research goals. Roth (1995) identified tests of models of individual choice and game theory (especially those involving industrial organization topics) as the early preoccupations of experimental economists. The later game-theoretic reframing, over the past dozen years, of nearly every field in economics – from microeconomic and industrial organization theory (e.g., Kreps 1990; Tirole 1988) to macroeconomic policy issues (Barro 1990) – provided a unifying theoretical framework that could easily be translated into experimental design.

Yet another aspect that helped to promote the comparative homogeneity of experimental practices within economics was its status as the "new kid on a hostile block" (Lopes 1994, p. 218). In light of severe criticisms from

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

prominent economists who claimed that it was impossible to make scientific progress by conducting experiments (e.g., Lipsey 1979; Russell & Wilkinson 1979; see *The Economist* May 8, 1999, p. 84), it is not surprising that economics was "more self-conscious about its science" (Lopes 1994, p. 218) and methodology than psychology. This explanation suggest that widely shared research goals and the prevalent rational-actor paradigm forced certain conventions and practices on experimental economists in a bid to gain acceptance within their profession. Last but not least it is noteworthy that throughout the 1970s and 1980s, experimental economics was concentrated at about a half dozen sites in the United States and Europe. We conjecture that this concentration helped the comparatively small number of experimental economists to agree on generally accepted rules of experimentation.

To conclude, several factors may account for the differing experimental practices in psychology and economics. Multiple research goals and the lack of a unifying theoretical framework that easily translates into experimental design may have promoted methodological variability in psychology. In contrast, the necessity to justify their practices within the discipline, an unusual concentration of key players in a few laboratories during the take-off phase, and the unifying framework provided by game theory may have helped economists to standardize their methodology.

6.3. Policy implication: Subject experimental practices to experimentation

As recently argued by Zwick et al. (1999, p. 6), methodological differences between psychology and economics are (at least partly) "derivatives" of differences in the assumptions commonly invoked (explicitly or implicitly) by economists and psychologists in the study of human choice. In our view, this argument must not be read as a justification to do business as usual. Neither psychologists nor economists have reason to avoid an interdisciplinary dialogue on the diverging methodologies for several reasons. First, some of the methodological differences – in particular, the (non)use of deception and scripts, but also the issue of abstract versus "natural" scripts (see note 4) – are not derivatives of theory differences; rather, they seem to be driven by methodological concerns that are largely independent of differences in theories (e.g., trust of potential participants).

Second, even those experimental practices that can be plausibly considered derivatives – for instance, financial incentives and repetition – can also be justified on the grounds of arguments not tightly linked with theory. For instance, it seems widely accepted that financial incentives reduce data variability (increase effect sizes and power of statistical tests; e.g., Camerer & Hogarth 1999; Smith & Walker 1993a). Similarly, a likely benefit of repetition is that participants have the chance to familiarize themselves with all the wrinkles of the unusual situation, and thus, their responses are likely to be more reliable (Binmore 1994).

Third, even if many psychologists do not endorse standard economic theory, they are often (particularly in recent decades) interested in testing its various assumptions (e.g., transitivity of preferences) or predictions. Those tests inevitably do entail the question of what is a "fair" test of standard economic theory – a question to which both psychologists and economists have to find a common answer. Finally, as economists move closer to psychologists' view of human choice – for instance, Simon's (1957) notion of bounded rationality, Selten's (1998) aspiration-adaptation theory, Roth and Erev's (1995) work on the role of reinforcement learning in games, Camerer and Ho's (1999) work on reinforcement and belief learning in games, Goeree and Holt's (in press a; in press b) incorporation of stochastic elements into game theory (see Rabin 1998, for many more examples) – one may envision a long-run convergence toward a common core of axioms in economics and psychology. A common ground concerning methodological practices – based upon an interdisciplinary dialogue and empirically informed design decisions – is likely to promote a theoretical convergence.

How can economists and psychologists establish such a common ground? As we pointed out earlier, we do not hold the conventions and practices in experimental economics to be the gold standard; they bring both benefits and costs. Nevertheless, there is a striking difference between the methodological approaches in psychology and economics: Economists seem to engage more often in cost-benefit analyses of methodological practices and to be more willing to enforce standards (e.g., to prohibit deception) if they are convinced that their benefits outweigh their costs. We suggest that psychologists, particularly in the context of justification, should also engage more frequently in such cost-benefit analyses and, as researchers, collaborators, and reviewers, enforce standards that are agreed upon as preferable. This is not to say that psychologists should adopt economists' practices lock, stock, and barrel. Rather, we advocate the subjection of methodological practices to systematic empirical (as well as theoretical) analysis. Applied to the variable of financial incentives, such an approach might be realized as follows (see also Camerer & Hogarth 1999).

Researchers seeking maximal performance ought to make a decision about appropriate incentives. This decision should be informed by the evidence available. If there is evidence in past research that incentives affect behavior meaningfully in a task identical to or similar to the one under consideration, then financial (or possibly other) incentives should be employed. If previous studies show that financial incentives do not matter, then not employing incentives can be justified on the basis of this evidence. In cases where there is no or only mixed evidence, we propose that researchers employ a simple "do-it-both-ways" rule. That is, we propose that the different realizations of the key variables discussed here, such as the use or non-use of financial incentives (or the use of different financial incentive schemes), be accorded the status of independent variables in the experiments. We agree with Camerer and Hogarth's (1999) argument that this practice would rapidly give rise to a database that would eventually enable experimenters from both fields to make data-driven decisions about how to realize key variables of experimental design.

This conditional do-it-both-ways policy should also be applied to two other variables of experimental design discussed here, namely, scripts and repetition of trials. In contrast, we propose that the *default* practice should be not to deceive participants, and individual experimenters should be required to justify the methodological necessity of each instance of deception to institutional review boards, referees, and editors. We do not exclude the possibility that there are important research question for which deception is truly unavoidable.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Nevertheless, we advocate a multi-method approach in which deception is replaced as much as possible by a collection of other procedures, including anonymity (which may undo social desirability effects; see the recent discussion on so-called double-blind treatments in research on dictator games, Hoffman et al. 1996), simulations (Kimmel 1996, pp. 108–13), and role playing (Kimmel 1996, pp. 113–16). We are aware that each of these methods has been or can be criticized (for a review of key arguments see Kimmel 1996). Moreover, it has been repeatedly pointed out that more research is needed to evaluate the merits of these alternatives (e.g., Diener & Crandall 1978; Kimmel 1996). A do-it-both-ways rule could be used to explore alternatives to deception by comparing the results obtained from previous deception studies to those obtained in alternative designs.

Let us conclude with two remarks on the APA rule of conduct concerning deception:

> Psychologists do not conduct a study involving deception unless they have determined that the use of deceptive techniques is justified by the study's prospective scientific, educational, or applied value and that equally effective alternative procedures that do not use deception are not feasible. (APA 1992, p. 1609)

Systematic search for alternative procedures – if enforced – may prove to be a powerful tool for reducing the use of deception in psychology. For instance, of the ten studies reported in Table 2, three used deception (Allison & Messick 1990, p. 200; Beeler & Hunton 1997, p. 83; Irwin et al. 1992, p. 111), including incorrect performance feedback, wrong claims about performance-contingent payments, and rigged randomization procedure. In our view, each of these deceptive practices was avoidable. Deception was also avoidable in another set of studies we reported here. In our sample of Bayesian reasoning studies (see sect. 3), we found that 37 out of 106 (35%) employed some form of deception (e.g., lying to participants about the nature of the materials used, falsely asserting that sampling was random, a precondition for the application of Bayes's theorem). If researchers met the APA requirement to seek alternatives to deception, they would have discovered "equally effective alternative procedures" already in the literature. Research in both psychology (e.g., Wallsten 1972; 1976) and economics (e.g., Grether 1980) shows that one can do without deception completely in research on Bayesian reasoning.

Finally, we propose (in concurrence with a suggestion made by Thomas Wallsten) that the assessment of the "prospective scientific value" of a study should not depend on whether or not a particular study can be conducted or a particular topic investigated. Rather, the question ought to be whether or not a theory under consideration can be investigated without the use of deception. This way, our assessment of the "prospective scientific value" of deception is closely linked to theoretical progress rather than to the feasibility of a particular study.

7. Conclusion

Some of the most serious (self-)criticism of psychology has been triggered by its cycles of conflicting results and conclusions, or more generally, its lack of cumulative progress relative to other sciences. For instance, at the end of the 1970s, Meehl (1978) famously lamented:

> It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics. (p. 807)

Since the 1970s, psychology's self-esteem has improved – with good reason. For instance, thanks to the increasing use of meta-analytic methods (Glass et al. 1981; Hedges & Olkin 1985), it has become clear that psychology's research findings are not as internally conflicted as once thought. As a result of this, some researchers in psychology have already called off the alarm (Hunter & Schmidt 1990; Schmidt 1992).

Despite this optimism, results in the "softer, wilder areas of our field," which, according to Rosenthal (1990, p. 775), include clinical, developmental, social, and parts of cognitive psychology, still seem "ephemeral and unreplicable" (p. 775). In his classic works on the statistical power of studies, Cohen (1962; 1988) pointed out two reasons (among others) why this is so. First, in an analysis of the 1960 volume of the *Journal of Abnormal and Social Psychology,* Cohen (1962) showed that if one assumes a medium effect size (corresponding to the Pearson correlation of .40), then experiments were designed in such a way that the researcher had less than a 50% chance of obtaining a significant result if there was a real effect (for more recent analyses, see Rossi 1990; Sedlmeier & Gigerenzer 1989). Second, Cohen (1988) suggested that many effects sought in various research areas in psychology are likely to be small. Whether or not one agrees with this assessment, the important point is that "effects are appraised against a background of random variation" (p. 13). Thus, "the control of various sources of variation through the use of improved research designs serves to increase effect size" (p. 13) and, for that matter, the power of statistical tests as well.

We believe that the realizations of the four key variables of experimental design in the areas of research discussed here contribute to the variability of empirical findings. Based on the evidence reviewed here, we argue that the practices of not providing a precisely defined script for participants to enact, not repeating experimental trials, and paying participants either a flat fee or granting a fixed amount of course credit only leave participants uncertain about the demand characteristics of the social situation "experiment." The fact that psychologists are (in)famous for deceiving participants is likely to magnify participants' uncertainty and second-guessing.

If our claim that a laissez-faire approach to experimentation invites lack of procedural regularity and variability of empirical findings is valid, and the resulting conflicting data indeed strangle theoretical advances at their roots (Loftus, in Bower 1997, p. 356), then discussion of the methodological issues addressed here promises high payoffs. We hope that this article will spur psychologists and economists to join in a spirited discussion of the benefits and costs of current experimental practices.

ACKNOWLEDGMENTS

We would like to thank Colin Camerer, Valerie M. Chase, Ward Edwards, Alexandra M. Freund, Gerd Gigerenzer, Adam Goodie, Glenn Harrison, Wolfgang Hell, John Hey, Charles Holt, Eva Jonas, Gary Klein, Martin Kusch, Dan Levin, Patricia Lindemann, Geoffrey Miller, Catrin Rode, Alvin Roth, Peter Sedlmeier, Tilman Slembeck, Vernon Smith, Ryan Tweney, Tom Wallsten,

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Elke Weber, David Weiss, and anonymous referees for many constructive comments. Special thanks are due to Valerie M. Chase and Anita Todd for improving the readability of our manuscript.

Correspondence should be addressed either to Ralph Hertwig, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany, or to Andreas Ortmann, Center for Economic Research and Graduate Education, Charles University, Politickych veznu 7, 11121 Prague 1, Czech Republic. Electronic mail may be sent to hertwig@mpib-berlin.mpg.de or to andreas.ortmann@cerge.cuni.cz.

NOTES

1. Sieber and Saks (1989) reported responses of 326 psychology departments. They found that of the 74% that reported having a participant pool, 93% recruited from introductory courses. The authors also found that "only 11% of departments have a subject pool that is voluntary in the strictest sense" (p. 1057). In contrast, economists recruit their participants in more or less randomly determined classes, through flyers or e-mail, often drawing on students from other disciplines. Because economists also typically use financial incentives, it is probably safe to assume that participation is voluntary.

2. For obvious reasons, we cannot reproduce the extensive instructions to participants here. However, we urge the reader who has not yet encountered a script-based study to take a look (e.g., pp. 1247 through 1253 in Camerer et al. 1989).

3. Most of the word problems listed here (e.g., conjunction task, engineer-lawyer task) are classic problems studied in the heuristics-and-biases program. Results and conclusions from this program have been hotly debated (for the different points of view, see the debate between Kahneman & Tversky 1996, and Gigerenzer 1996).

4. Scripts may be content-free or enriched with social context. In an attempt to control home-grown priors (i.e., beliefs and attitudes that participants bring to the experiment), the scripts provided by economists are typically as content-free as possible. From the perspective of the experimenter, such environments may be precisely defined, but they seem to tax the cognitive abilities of participants more than seemingly more complex but familiar real-world scripts, because they take away the "natural" cues that allow participants in real-world environments to understand situations. Assuming the existence of domain-specific reasoning modules, Cosmides and Tooby (1996) even argue that the starkness of laboratory environments prevents specialized inference engines from being activated, and that mismatches between cues and problem types are far more likely under artificial experimental conditions than under natural conditions. This trade-off between control of home-grown priors and accessibility of "natural" cues has long been discussed in psychology (e.g., Bruce 1985; Koriat & Goldsmith 1996 for the real-life/laboratory controversy in memory research; see Goldstein & Weber 1997 for the issue of domain specificity in decision making, and Winkler & Murphy 1973 for their critique of the bookbag-and-poker chips problem in research on Bayesian reasoning). It has also recently been addressed in studies by economists (e.g., Dyer & Kagel 1996; Schotter et al. 1996).

5. Harrison (1989; 1992) argued that many experiments in economics that provide financial incentives dependent on performance nevertheless lack "payoff dominance." Lack of payoff dominance describes essentially flat maxima, which make it relatively inexpensive for participants not to choose the theoretically optimal action (von Winterfeldt & Edwards 1982). The implication of Harrison's critique is that performance in a task can only be classified as "irrational," "inconsistent," or "bounded" if the difference between the payoff for participants' actual behavior and that for optimal behavior in an experiment is monetarily significant to participants given their standard hourly wage. "Significant" could mean, for example, that the potential payoff lost owing to nonoptimal behavior in a one-hour experiment exceeds one hour's worth of wages for the participant and 25% of total payoffs obtainable. If the difference between the payoff for the participant's actual behavior and that for optimal behavior is, say, only 5%, one could argue that the payoff decrement participants accept by not behaving optimally is too trivial to be considered "irrational."

6. The systematic study of financial incentives can help us question long-held beliefs. For instance, Koriat and Goldsmith (1994) reported that memory accuracy (i.e., the percentage of items that are correctly recalled) is strategically regulated, that is, "subjects can substantially boost their memory accuracy in response to increased accuracy motivation" (p. 307). Koriat and Goldsmith stressed that their results contrast sharply with the general observation from quantity-oriented research that people cannot improve their memory-quantity performance when given incentives to do so. Participants in a high-accuracy-incentive condition were more accurate than those in a moderate-accuracy-incentive condition (*eta* = .58, a large effect according to Cohen 1988; calculated from data in Koriat and Goldsmith's 1994 Table 3).

7. Needless to say, the implementation of financial incentives has its own risks. It is, for example, important to ensure that payments are given privately. As a referee correctly pointed out, public payment can be "akin to an announcement of poor test performance and might violate a number of ethical (and, in America, perhaps legal) standards, and is all the more likely to negatively impact mood." Private payment is the standard practice in economics experiments.

8. One reviewer referred us to Frey's (1997) discussion of the hidden costs of extrinsic rewards. Frey's book, as thought provoking and insightful as it often is, takes as its point of departure the same literature that Eisenberger and Cameron (1996) discussed and took issue with. As mentioned, we agree that money does not always work as a motivator, but we believe that more often than not it does. Let us consider Frey's example of professors. Professors who are so engaged in their profession that they teach more than the required hours per week may indeed react with indignation when administrators try to link remuneration more closely to performance and therefore reduce their extra effort. There are, however, also professors who "shirk" (the term used in principal-agent theory) their teaching obligations to do research, consulting, and so forth. In fact, shirking has been identified as the major driver of the inefficiency of educational institutions in the United States (Massy & Zemsky 1994; Ortmann & Squire 2000). While consulting has immediate material payoffs, at most institutions research translates into higher salaries and, possibly more important, payoffs such as the adulation of peers at conferences (Lodge 1995). It is noteworthy that the activities that professors engage in involve by their very nature self-determination, self-esteem, and expression possibility and therefore are particularly susceptible to "crowding out." In contrast, most laboratory tasks do not prominently feature these characteristics.

9. What constitutes deception is not easy to define (see Baumrind 1979; Rosenthal & Rosnow 1991). Economists seem to make the following pragmatic distinction, which we endorse: Telling participants wrong things is deception. Conveying false information to participants, however, is different from not explicitly telling participants the purpose of an experiment, which is not considered deception by either economists (McDaniel & Starmer 1998; Hey 1998) or psychologists known to be opposed to deception (e.g., Baumrind 1985). However, to the extent that absence of full disclosure of the purpose of an experiment violates participants' default assumptions, it can mislead them, and therefore should be avoided.

10. To avoid many "I don't know" responses, we asked economists to estimate how often psychologists in general (rather than researchers in JDM) use various practices. This may explain why their estimates for the use of deception were so high.

11. There are also regulatory standards in psychology – possibly the best examples are the treatment group experiments and null-hypothesis testing (see Danziger 1990). Null-hypothesis testing was, and to a large extent remains, a self-imposed requirement

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

in psychology despite continuous controversy about its use. How is null-hypothesis testing different from the key variables of experimental design considered here? Gigerenzer and Murray (1987) argued that "the inference revolution unified psychology by prescribing a common method, in the absence of a common theoretical perspective" (p. 22). One may speculate that null-hypothesis testing still predominates in psychology because abandoning it may be perceived as abandoning the unification of psychological methodology. The key variables of experimental design considered in this article have never filled this role.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Purposes and methods

Jonathan Baron

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196. **baron@cattell.psych.upenn.edu**
www.sas.upenn.edu/~jbaron

Abstract: The methods of experiments in the social sciences should depend on their purposes. To support this claim, I attempt to state some general principles relating method to purpose for three of the issues addressed. (I do not understand what is not a script, so I will omit that issue.) I illustrate my outline with examples from psychological research on judgment and decision making (JDM).

Repetition

(1) This is useful when we want to study practice effects.

(2) It is useful when we want to get as much data from each subject as possible, because our marginal cost of collecting them is lower than that of additional subjects.

(3) It is useful when we want to fit mathematical models to each subject's data.

(4) It is useful when we are concerned about subjects' understanding the tasks and feel that experience will help them understand. Note that we can try to insure understanding in other ways (which can be combined): incentives, test questions, reading instructions aloud (so that subjects cannot skip them), using familiar examples, and pretesting the instructions themselves.

(5) It is useful when we want to model real-world situations involving repeated decisions or judgments of the same type, such as those made by physicians most of the time, but not those made by most patients.

On the other hand, many decisions that people make are not repeated, and people must make them by applying general principles learned elsewhere to a case that is somewhat different from those on which they learned. Some interpretations of economic theories imply that the principles of rational choice have been learned in a general form, so they ought to apply everywhere. Indeed, this is part of the implicit justfication of using the laboratory to simulate the world outside of it. If a principle is generally true of human behavior, then it ought to apply everywhere, including unfamiliar situations.

(6) Repetition is often avoided in the study of experimental games, because repeating the game could change its nature by introducing sequential strategies. Often this problem is avoided by convincing the subjects that they are playing with different players in each game.

(7) Another reason not to use repetition is that accumulated gains or losses may affect subsequent decisions. This problem may be avoided by telling subjects that only one trial will count.

Incentives and contingent payoffs

(1) These are useful when we want to study responsiveness to the incentives themselves, to test some hypotheses derived from economic theory or learning theory.

(2) Incentives are also useful when the response may be affected by social desirability, as in the case of cooperation in social dilemmas. Most of the psychological study of social dilemmas uses incentives for just this reason.

(3) A closely related problem is the use of payoffs in contingent valuation experiments. People may want to pay more for a good than they would acually pay, and, as the target article points out, stated WTP is often greater than actual WTP.

(4) It is not obvious, though, which answer is correct. People may express true values in what they say they are willing to pay and then regret their actual failure to contribute.

(5) It is impossible to use real payoffs when the experimenter does not control the good being evaluated.

(6) Payoffs may help subjects try hard when the task is very difficult.

(7) Payoffs are expensive. Hence they reduce the amount of research that can be done with a fixed budget.

(8) Payoffs are unnecessary when the research concerns a well-replicated effect, known to be found in situations of interest. A great deal of JDM research concerns manipulations of such effects.

(9) Payoffs are useless when there is no right answer. This is typically true in studies of judgment, as opposed to decision making. Studies of judgment are, of course, part of the set of studies of "judgment and decision making." These studies also include judgments of fairness and other moral judgments. Even if the experimenters think they know the right answers to moral questions, it is inappropriate for us to pay subject to give them.

Deception

(1) Deception is harmful because it creates externalities. Deception makes it more difficult for future experimenters to induce subjects to believe their stories. Experimenters may avoid this by establishing a reputation, perhaps at the level of the laboratory or even the discipline (such as experimental economics), or by offering a warranty in the form of payment to any subject who discovers deception, as established by a neutral third party.

Of course, very few experimenters, in any field, tell their subjects the hypothesis of interest. This is deception by omission. Subjects are used to this. But it is, in fact, an externality. It creates a belief in subjects that the experimenter usually withholds the purpose of the study, making it more difficult for experimenters to tell subjects the purpose (even in general terms) when it is in fact a good thing for the subjects to know.

(2) Despite its costs, deception has benefits in terms of resource savings. Statistical power is increased in a game experiment if everyone plays with (or against) the same computer program rather than (as they are told) other people.

(3) Deception is sometimes the only way to study the question of interest, as when emotions must be induced. Induction of emotions requires the use of a theatrical display, carefully set up. This sort of use of deception is rare in JDM. One possible example is the use of positive rewards like candy to examine the effects of mood. This is deceptive because the subjects are not told, "We are giving you this candy to make you feel good."

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Financial incentives do not pave the road to good experimentation

Tilmann Betsch and Susanne Haberstroh

Psychological Institute, University of Heidelberg, D-69117 Heidelberg, Germany. **tilmann.betsch@urz.uni-hd.de**
susanne_haberstroh@psi-sv2.psi.uni-heidelberg.de
www.psychologie.uni-heidelberg.de/AE/sozps/tb/TB_home.html
www.psychologie.uni-heidelberg.de/AE/sozps/SHengl.html

Abstract: Hertwig and Ortmann suggest paying participants contingent upon performance in order to increase the thoroughness they devote to a decision task. We argue that monetary incentives can yield a number of unintended effects including distortions of the subjective representation of the task and impaired performance. Therefore, we conclude that performance-contingent payment should not be generally employed in judgment and decision research.

It is a common practice in experimental economics to employ financial incentives to increase performance in decision making. The conviction behind this policy is that performance-contingent payment motivates participants to deliberate more thoroughly on the task. H&O suggest adopting this principle more often in psychological experimentation. We doubt that such a practice is generally worth the effort.

Payment may increase performance only in those tasks which obey the following principles: (1) External criteria of performance must be exactly and objectively determined. (2) External and internal criteria must match. For example, in a decision problem, the subjective importance of goals should not differ from their ranking according to external standards. (3) Elements and structure of a given problem must be well-defined and maintained in subjective representation. (4) A deliberate strategy must be the most appropriate means to solve the problem. This might be the case, for example, when the set of given information can be overlooked, and when time and cognitive resources are available for sequential processing.

In the remainder of the paper, we argue that it is difficult to determine external criteria in most of the decision domains studied by psychologists, and that the other principles are generally likely to be violated, due to fundamental properties of the human processing system.

Lack of exact criteria. Defining external criteria is a necessary antecedent of performance-contingent payment. In some decision tasks (e.g., monetary gambles), it is easy to determine external criteria because alternatives pay off only on one dimension (money), and theories are available that set-up normative standards (utility theory). Yet many domains of judgment and decision making studied by psychologists involve multiple-attribute decisions (e.g., selection among partners, political candidates, consumer products, social choices). Distinguishing good from bad choices in such domains would require us to be able to determine objectively the *importance* of the attributes involved (e.g., a partner's physical attractiveness, age, health, wealth). Needless to say, this is difficult, if not impossible, for many real world decision problems.

Discrepancies between external and internal criteria. Sometimes it might be possible to approximate objective standards of performance, even in non-monetary, multi-attribute judgment tasks; albeit people's internal standards will not necessarily match the external ones. Mere thinking can blur the match between criteria. Using a verbal protocol methodology in framing studies, Maule (1989) found that people tend to evaluate the outcomes of the given alternatives on a couple of dimensions, which are not explicitly contained in the task (e.g., norm and moral). Similar observations have been made by Kahneman et al. (1987), and Igou et al. (1999). Financial incentives can even foster the mismatch between internal and external criteria because they encourage people to think about the problem. This can lead to the consideration of additional criteria, which otherwise would have been viewed irrelevant for the task. As a consequence, the quality of performance can decrease (Wilson & Schooler 1991).

Deviations in problem representation. The structure of given problems is generally unlikely to be maintained in subjective representation, because input information is always processed in the light of prior knowledge (Bruner 1957). Prior knowledge effects are well-documented in the judgment and decision literature. For example, research on the conjunction fallacy (e.g., the Linda problem) has provided ample evidence indicating that judgments reflect subjective constructions of the problem, which systematically deviate from the surface structure of the given task (e.g., Betsch & Fiedler 1999; Dulany & Hilton 1991). A plethora of prior knowledge effects have been accumulated in research on recurrent decision making (Betsch et al. 2000). Most importantly, prior knowledge can overrule given information and can lead to maladaptive choices, even when the individual has enough time to carefully deliberate on the task (Betsch et al., in press). Generally, the more intensely people think about stimuli, the higher the likelihood that prior knowledge effects become more pronounced. Consequently, financial incentives will foster prior-belief biases in judgments and decisions.

Automatic modes can outperform deliberative modes of processing. The notion that intense thinking increases performance ignores more than a hundred years of psychological research on automaticity and implicit cognition. To say it pointedly, conscious processes are not necessary for perception, learning, memory, language, and adaptive behavior (Jaynes 1976). Although thinking is functional under certain conditions (Baumeister & Sommer 1997), it plays a subordinate role in most of the activities of everyday life (Bargh 1996). Today, psychology witnesses a renaissance of research on automaticity (Wyer 1997). In contrast to deliberative modes, which involve serial processing, information can be processed in parallel under automatic modes. Thus, automaticity allows for holistic assessment of complex problems at a rapid pace. Moreover, the major part of implicit information in memory is accessed and processed in automatic modes. If implicit knowledge provides a representative sample of task relevant information, automatic or intuitive processes can yield a remarkable degree of accuracy in judgments and decisions. In a couple of experiments, Betsch and colleagues had participants encode a large amount of information about share performance on several stock markets over a period of time (Betsch et al. 2001). Retrospective intuitive judgments about shares' performance were highly accurate. Conversely, thinking decreased in accuracy. In deliberative mode, participants attempted to access concrete memories, and consequently, judgments were biased toward the most accessible ones. Similar results were obtained in the domain of frequency judgment (Haberstroh et al. 2000). Participants estimated frequencies of their own behaviors after having repeated these behaviors in a series of preceding decision tasks. Half the participants gave intuitive judgments, the other half was asked to think carefully. Moreover, they expected performance-contingent payment. Again, intuitive judgments were remarkably accurate, whereas thinking led to severe biases in judgment reflecting the availability of examples of the behaviors.

The above examples show that researchers should be careful in using financial incentives in order to increase performance. Even if standards of performance can be objectively determined, financial incentives might have counterproductive effects. Incentives can encourage the individual to change the representation of a given task. Moreover, they might impede participants, exhausting the full potential of their processing system by fostering deliberative modes of processing, which can produce inferior results compared to intuitive processing modes. Therefore, we conclude that financial incentives do not pave the road to good experimentation.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Typological thinking, statistical significance, and the methodological divergence of experimental psychology and economics

Charles F. Blaich[a] and Humberto Barreto[b]

[a]*Department of Psychology;* [b]*Department of Economics, Wabash College, Crawfordsville, IN 47933.* **{barretoh; blaichc}@wabash.edu**
www.wabash.edu/depart/economic/barretoh/barretoh.html

Abstract: While correctly describing the differences in current practices between experimental psychologists and economists, Hertwig and Ortmann do not provide a compelling explanation for these differences. Our explanation focuses on the fact that psychologists view the world as composed of categories and types. This discrete organizational scheme results in merely testing nulls and wider variation in observed practices in experimental psychology.

We agree with Hertwig and Ortmann's (H&O's) description of current practices in experimental economics and psychology. They make convincing arguments that scripting, incentives, repeated trials, and the lack of deception act to increase the consistency of findings in experimental economics. We disagree, however, with the explanation for the variation in the observed practices of the two disciplines. H&O quickly mention several factors (see sect. 6.2, "Why do the methodological practices differ?") which may be contributing, but not fundamental, answers to the question. We believe there is a sharp methodological divide between psychologists and economists, which then drives the differences in their respective experimental practices.

Budding experimental psychologists and economists are both schooled in the ways of the null and alternative hypotheses, t-tests, and Type 1 and 2 errors. Despite this common background, experimental psychologists and economists do not typically emphasize the same questions when they analyze their studies. Experimental psychologists generally place a great deal of importance on rejecting the null hypothesis, while experimental economists are more likely to emphasize both rejecting the null hypothesis *and* estimating the magnitude of parameters. The fact that experimental psychologists tend to assign much more importance to rejecting the null hypothesis but less importance on making precise parameter estimates than experimental economists plays an important role, in our view, in creating the differences in the two fields that the authors describe.

The null and alternative hypotheses refer to two mutually exclusive and exhaustive states of affairs. The null usually posits that an independent variable has no effect and any patterns in the data are due solely to chance. The alternative is not a specific hypothesis, but rather, the assertion that the null hypothesis is false. By designing and analyzing studies with the sole purpose of rejecting the null hypothesis, experimental psychologists avoid the burden of deriving, testing, and replicating specific predictions by a theory of how variables should be related.

Consider the typical approach that an experimental psychologist might use to examine the relationship between motivation and academic performance. The psychologist may have a vague hypothesis that "students who are more motivated will get higher grades." Generally, there would be no theory which specifies the magnitude of the effect of motivation nor the form of the relationship between motivation and academic performance. Instead, the question would simply be, "Does motivation have an effect on academic performance, yes or no?" This question, in turn, would boil down to the question of whether the effects of motivation on academic performance are statistically significant.

Given the binary nature of the research question and the lack of a theory that creates specific predictions, it would not be surprising to find a wide variety of experimental designs that would be employed to test the hypothesis, and a wide range of "statistically significant" outcomes that would all be seen as consistent despite substantial variation in measured effects.

But pointing to the higher propensity of psychologists to test nulls begs the question of why experimental psychologists are more likely to design research that is aimed at answering yes/no questions. We believe that a fundamental reason is the emphasis in experimental psychology on types and categories. This emphasis is reflected and perpetuated by their continued reliance on ANOVA and factorial research designs that were pioneered by Fisher (1956). To continue our research example on motivation and academic performance, since their goal is simply to determine if motivation matters, most experimental psychologists would manipulate motivation so that one or two groups received a treatment that would alter their level of motivation, and they would probably add a no-treatment control group. While it is easy to imagine that a good experimental psychologist could create a more complicated design that would include more groups, it is hard to imagine an experimental psychologist who would consider manipulating motivation continuously. Yet, by transforming and truncating a continuous variable such as motivation into a discrete two- or three-level variable, experimental psychologists lose information about the magnitude and form of the relationships that involve that variable.

Contrast the psychologist's categorical view of the world with that of the economist. The pillars of economics, supply and demand, are continuous functions that describe how much more is offered by producers and how much less is desired by consumers as price rises. Economists do not ask if education affects incomes; they attempt to estimate the rate of return to an additional year of education. Of course, economics cannot match the usual case in the natural sciences, where a continuous variable is embedded in a prediction with a specific functional form.

Given the categorical view of the world that is facilitated by their designs and statistical analyses, there is no reason for psychologists to develop theories with more specific predictive power. Indeed, there is a greater risk in doing so. As Meehl (1978) pointed out, it is far easier to falsify a theory which makes a specific prediction about the form and magnitude of the relationship between variables. With null hypothesis testing as a guide, research on $F=ma$ would be reduced to a null hypothesis that there was no relationship between force, mass, and acceleration; an alternative hypothesis that there was some relationship between these variables; and a design which included "force" and "no-force" groups. Any theory that predicted some relationship between these entities (assuming sufficient power) would be supported by the data, and there would no reason to adopt a consistent set of methods that could reliably reproduce a specific parameter effect.

H&O have correctly described the differences in experimental practices in economics and psychology. Our contribution to their excellent article is to highlight the role of null hypothesis testing in experimental psychology, which we believe is ultimately due to psychology's focus on types and categories, instead of continuous functions. When psychology changes its world view from rigid, discrete categories to continuous, gradual movements, testing nulls will be replaced by estimating parameters and experimental practices will increasingly converge to a standard.

Economic and psychological experimental methodology: Separating the wheat from the chaff

Hasker P. Davis and Robert L. Durham

Department of Psychology, University of Colorado, Colorado Springs, CO 80933. **{hdavis; rdurham}@brain.uccs.edu**
www.web.uccs.edu/{hdavis; rdurham}

Abstract: Hertwig and Ortmann suggest methodological practices from economics (script enactment, repeated measures, performance based payments, and absence of deception) for psychology. Such prescriptive

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

methodologies may be unrepresentative of real world behaviors because people are not: always behaving with complete information, monetarily rewarded for important activities, repeating tasks to perfection, aware of all contributing variables. These proscriptions, while useful in economics, may obfuscate important psychological phenomena.

Hertwig and Ortmann (H&O) advocate a fuller usage of four methodological procedures from economics for psychological research. These procedures are "script enactment, repeated trials, performance-based payments, and a proscription against deception." In economics, script usage reduces participant uncertainty, repeated measures allows the participant to gain familiarity with the test procedure, financial incentive clarify performance demands, and the proscription of deception makes the experimental situation more predictable. These procedures either currently receive adequate consideration in psychological research, or their implementation as suggested by H&O will not facilitate addressing research questions in psychology.

H&O contend that participants who are naive cannot perform tasks as well as participants who are entirely trained and knowledgeable about the task. However, people are continuously interacting with the world without complete knowledge of the processes underlying their behaviors/thoughts. As the economist Simon (1979) has put it, people are content to be "satisficed" [sic]. As a matter of fact, most decisions are made on a day-to-day basis, even in economic and fiscal areas, with incomplete knowledge (e.g., just look at USA and UK politicians!). Persons in the "real world" rarely are informed to the extent that they could rehearse scripts in their problem solving and decision making activities. While it is useful to model those activities to observe maximum and differential effects of various techniques, to argue that those effects represent how people actually behave naturally is naive in itself. For example, one might expect judges and attorneys to be more informed mock jurors than undergraduates, but in experimental situations the two groups reach identical verdicts (for review, see Bornstein 1999).

H&O argue that paid participants are more highly motivated to perform well than unpaid participants who receive academic credit. First, it is completely understandable that researchers concerned with economic variables should be involved with the monetary payment of participants and the differential effects of those payments on performance. However, and again, in the "real world" people engage in all sorts of behaviors that are not monetarily relevant (e.g., memory tasks, problem solving, social interactions, child rearing, jury decision making, etc.). To always investigate such behaviors using money would be inappropriate, and not valid. Second, and as noted by H&O, social psychologists and personality theorists have already investigated the differential effects of different reward systems and values (intrinsic vs. extrinsic) on various behaviors or traits in order to determine their differential effects (e.g., Bergin 1991; Deci & Ryan 1987). They do, in fact, exist under some circumstances and not in others. To suggest that the study of the myriad of human activities is best exemplified by a single reward system is not a viable approach.

H&O suggest that a major advantage of the within design is the participants' enhanced familiarity with the test procedure. Other advantages of the design include the requirement for few participants, a lesser time requirement, reduced error variance, and a more powerful test of the variable of interest (i.e., decreased likelihood of Type II error). H&O are correct to an extent in their arguments about using repeated measures designs rather than single snapshot between designs. However, they neglect several deficits of this design that keep it from always being the design of choice. H&O note the social nature of the relationship between subject and experimenter and offer their methodological recommendations as means to reduce the uncertainty of the experimental social situation. Their recommendations do change the experimental situation, but not always in the desired way. Repeated measures can make the demand characteristics of the experiment more apparent to the subject and introduce a confound by changing his/her performance. Other well-known problems associated with a repeated measures design are practice effects, boredom effects, fatigue effects, and carryover effects. For example, in the case of carryover effects, if rates of forgetting are being examined, once a subject has participated in one delay condition, their performance in other delay conditions may be affected by their expectation of a delayed retention test. Grice (1966), in contrasting the advantages and disadvantages of a within design to a between design, states: "[s]uch an experiment may be good or bad, but if the experimenter thinks that he has merely done more efficiently the same investigation as the independent group experiment, he is mistaken" (p. 488).

Some of the insights into human behavior gained from research using deception would most likely not have been easily obtained using alternative methodology. For example, in Milgram's (1963) study of obedience to authority, the extent to which individuals complied was much greater than virtually anyone anticipated. Despite numerous ethical criticisms of this study, numerous follow up studies have forced an unpleasant awareness of our capacity for brutish behavior. However, it appears that H&O are not basing their arguments against deception so much on ethical concerns, as on concerns about contamination of the participant's development of suspicion and potential behavioral changes in studies not using deception. Since H&O offer no viable alternative to deception, one can only say psychologists have been conscientious in their search for alternatives to deception when the psychological processes under investigation warrant or allow options.

We certainly agree with H&O that psychology could benefit from some improvements in the practice of methodological rigor. However, we do not think psychologists need to become economists to do so. It is really the question being asked that will always determine the methodology. If we get caught up in prescriptions for rigorous formal research methodology, what becomes of the participant-observer methodology of cultural anthropology or single case studies in clinical psychology, both of which have enriched our understanding of human behavior? The areas of economics and psychology and their questions of interest seem different enough that different methodologies and goals can be used to enhance understanding of behavior. Indeed, a proscription of research implied by H&O seems to verge on a call for the application of a single method ethnocentrism for major areas of human behavior. This seems unproductive, and is, of course, impractical. Similarly, H&O's suggestion to do it "both ways" is impractical because of fiscal and time requirements. Scientists in economics and psychology will struggle to find the methodology that best answers the question they are addressing. The consumer of the literature must then separate the wheat from the chaff.

On accumulation of information and model selection

Ido Erev

Columbia Business School, New York, NY 10027; Industrial Engineering and Management, Technion, Haifa 32000, Israel. **erev@tx.technion.ac.il**

Abstract: This commentary extends Hertwig & Ortmann's analysis by asking how stricter model selection conventions can facilitate the accumulation of information from experimental studies. In many cases researchers are currently motivated to summarize their data with ambiguous and/or multi parameter models. A "generality first" convention can help eliminate this problem.

Hertwig and Ortmann (H&O) convincingly argue that the wide range of acceptable experimental procedures used in psychology increases data variability and can impair accumulation of knowledge. To eliminate this problem they suggest the use of the stricter conventions used in experimental economics. The "do it both

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ways" convention they suggest is expected to reduce noise and for that reason to facilitate accumulation of knowledge.

Since H&O limit their discussion to experimental conventions, they did not consider the possibility that a change in model selection conventions can also help solve the problem they describe. The main goal of the current commentary is to discuss the value of this option and its interaction with the suggestion made by H&O. It is argued that stricter demand on the development of descriptive models can help accumulation of knowledge and is relatively robust to noise in the data.

Like the stricter experimental conventions proposed by H&O, the model selection convention discussed here is popular in economics. The basic idea is that a counter example that implies that a descriptive model is inaccurate does not imply that the model should be replaced. Because descriptive models are only approximations, a discovery of a situation in which they are violated has limited implications. Only when an alternative model is proposed which provides a more accurate summary of the new data as well as the data captured by the original model should a model be replaced. This convention is referred to here as the "generality first" rule.

The generality first rule is most effective for models that make *ex ante* quantitative prediction of behavior. Models of this type are assumed to have robust parameters in a wide set of tasks. Thus, their parameters can be derived on one subset of tasks and then evaluated on a second set of experiments. In this case the generality first convention seems natural. For example, if 80 experiments show that a particular model provides a useful prediction of behavior in a well-defined set of tasks, it should be clear that a rejection of the model in Experiment number 81 does not imply that the model is not useful. The model should be replaced only if an alternative model is proposed which provides a better summary of the 81 tasks.

Thus, if the results of Experiment number 81 are not reliable, they are unlikely to impair accumulation of knowledge. As long as we do not overfit the data, attempting to model the noisy results together with the accumulated data set is expected to reveal that the best general model (summary of the large data set) does not change.

Given the potential value of the generalization first rule in facilitating accumulation of knowledge, it is surprising to see how little this rule is used in psychology and in experimental economics. I believe that the main reason for this limited usage is a problematic incentive structure. My colleague Daniel Gopher says (following a similar argument by Malcolm Ritchie): "the citation maximizing strategy in Psychology is to start with an extreme theoretical assertion (to get attention) and then to add some ambiguity to insure that the model will not be easily rejected." Ambiguity can be added by using natural language (rather than a quantitative model) or by developing a model with multisituation-specific parameters (that can fit every data set but predict very little). The combination of the two ambiguity generating approaches is, of course, even more "efficient."

Unfortunately, this problematic incentive structure implies that enforcing a "generalization first" rule is not easy. In fact, the popularity of multi-parameter ambiguous models is increasing even in experimental economics. In Roth et al. (2000) we try to address this problem by suggesting an explicit calculation of the predictive value of descriptive models. We hope that the availability of explicit measure of predictive value will help reduce the incentives to propose ambiguous models.

Interesting to note, our suggestion and the "do it both ways" convention suggested by Hertwig and Ortmann are expected to support each other. Practice of the "do it both ways" rule should increase the availability of more standardized data that will reduce the attractiveness of situation specific parameters. And an increase in the popularity of general models will reinforce experimenters whose manipulations are clear enough and can be modeled.

Behavioral and economic approaches to decision making: A common ground

Edmund Fantino and Stephanie Stolarz-Fantino

Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. **{efantino; sfantino}@ucsd.edu**

Abstract: Experimental psychologists in the learning tradition stress the importance of three of the authors' four key variables of experimental design. We review research investigating the roles played by these variables in studies of choice from our laboratory. Supporting the authors' claims, these studies show that the effects of these variables are not fixed and should not be taken for granted.

Hertwig and Ortmann make an important contribution by identifying four key variables of experimental design that tend to be treated very differently by economists and psychologists. While the authors specifically exclude experiments in learning from their criticisms of research practices, some of their most interesting points can be supported, and qualified, by citing examples of relevant research from experimental psychology in the learning tradition.

A key design variable in behavioral research is the use of repeated rather than single trials. Much of our own research on decision making has involved repeated trials. In contemplating base-rate neglect, we wondered if the effect would rapidly dissipate if subjects were presented with repeated trials, especially under conditions likely to assure attentive, motivated subjects (Stolarz-Fantino & Fantino 1990). In a series of experiments Goodie and Fantino utilized a matching-to-sample (MTS) procedure to mimic crucial aspects of one of the classic versions of the base-rate problem. From a behavioral perspective, base-rate problems such as the taxicab problem are examples of multiple stimulus control: control by the sample (or "witness") cue and control by the base rates ("cab color frequencies" or probabilities of reinforcement for choosing either alternative, independent of the sample cue). The MTS procedure allows us to manipulate separately these two sources of stimulus control with repeated trials in a behavioral setting. Subjects could be presented with up to 400 trials in an hour. Using this procedure we could readily manipulate the base-rates and sample accuracy and assess the control of decisions by sample accuracy and base rates. In all of our work the control by sample accuracy overwhelms control of base rates, leading to non-optimal choices. Moreover this behavioral base-rate neglect persists for several hundred trials (e.g., Goodie & Fantino 1995; 1996; 1999b). In one experiment, in which we assessed choice over 1,600 trials (Goodie & Fantino 1999a), base-rate neglect was eventually eliminated. However, the eventual disappearance of base-rate neglect in no way minimizes its importance: Life rarely offers 1,600 trials (Fantino 1998).

A second factor stressed by the authors is the use of financial incentives. In cases where there is no a priori reason to assume that financial incentives matter the authors propose that a "do-it-both-ways" rule be employed. In our research on the MTS analog of the base-rate problem we found no significant effect of financial incentives (Goodie & Fantino 1995). Thus, we generally omit them. Similarly, in a research program assessing the reinforcing effectiveness of information, we have found comparable results whether or not the points that our subjects earned were backed up by money (Case & Fantino 1989). This by no means suggests that the size of the incentive is generally irrelevant and the authors are right to question studies that assume otherwise in the absence of data. Indeed this may be an important issue in research on self-control in the behavioral laboratory (Fantino 1966; Logue 1988; Rachlin 1995). While many studies of self-control with pigeons show extreme temporal discounting of delayed rewards, the hallmark of impulsive decision-making, it is much more difficult to demonstrate impulsive behavior in humans. This is inconsistent with anecdotal evidence from everyday life suggesting that humans have great difficulty exercising self-control. One possibility

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

is that the incentives offered humans in self-control studies pale beside the incentives offered pigeons (typically maintained at 80% of their free-feeding body weights and under a 23-hr food-deprivation regimen). In any event there is little question that, under some circumstances, financial (and other strong) incentives may greatly affect decisions. As the authors conclude it is important to identify the conditions under which this is so.

We also agree with the authors that what subjects are told about a task can be of central importance, even when deception is not involved. Moreover in our experience it can be difficult to anticipate the effects of instructions. For example, Case et al. (1999) instructed subjects about the MTS task by conveying with simple "picture instructions" the random nature of the correct responses and the exact base rate that they would be experiencing. This manipulation had no effect, even though subjects were required to accurately count the 100 outcomes of a sequence of outcomes "generated in exactly the same way that the computer will generate the sequence of correct alternatives in your sessions" (Case et al. 1999, p. 324). Thus, instructions do not necessarily affect performance in the manner expected.

The effects of instructions no doubt interact with subjects' histories. Arkes and Ayton (1999) and Goodie and Fantino (1996) have argued that non-optimal decision effects such as the sunk-cost effect and base-rate neglect may result from preexisting (learned) associations. In non-humans such lapses in decision making are uncommon. For example, Hartl and Fantino (1996) report an experiment with pigeons that employed a procedure comparable to that of Goodie and Fantino (1995) with humans. Whereas Goodie and Fantino found base-rate neglect over hundreds of repeated trials, even with monetary incentives for correct responses, Hartl and Fantino's pigeons performed optimally in all conditions. In research on persistence of commitment, Sonia Goltz has shown that humans with a history of variable reinforcement are much more persistent in pursuing a non-optimal decision path than those with a more regular history of payoffs (e.g., Goltz 1993; 1999). When viewed in a historical context, our decisions may not be seen as more rational but at least their etiology may be better understood.

In research on the conjunction fallacy, using a conventional story format, we have looked at the effects of repeated trials, monetary incentives, and feedback (Stolarz-Fantino et al., unpublished; Zizzo et al. 2000). We have not found improvement over 6 repeated trials; the fallacy has also remained robust to hints, feedback, and payment for correct answers. We are currently collecting data on story versions of the conjunction and base-rate problems using a larger number of repeated trials and comparing feedback and no-feedback conditions. We agree with the authors that it is advantageous to "do it both ways."

Are we losing control?

Gerd Gigerenzer

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, D-14195 Berlin, Germany. **giger@mpib-berlin.mpg.de**
www.mpib-berlin.mpg.de/Abc/home-d.html

Abstract: Most students are trained in using but not in actively choosing a research methodology. I support Hertwig and Ortmann's call for more rationality in the use of methodology. I comment on additional practices that sacrifice experimental control to the experimenter's convenience, and on the strange fact that such laissez-faire attitudes *and* rigid intolerance actually co-exist in psychological research programs.

Methodological practices are rarely the subject of reflection. Most of us have not chosen a practice; someone else did this for us. Yet we tend to defend what we received, following habit, group loyalty, and peer pressure. Hertwig and Ortmann (H&O) do a great service in challenging this proclivity. They ask us to reflect on experimental practices. One might object that differences in practice directly mirror differences in the subject matters of psychology and economics – just as it is natural to use microscopes for viewing things tiny and near but telescopes for ones large and distant. One might thus conclude: Leave the psychologists in peace and let the economists do what they do. My first point is that this "naturalistic" argument is invalid: experimental practices, and their enforcement, in fact vary strikingly within psychology and often resemble those in economics.

What Wundt and Skinner have in common. In Wundt's laboratory, known as the first psychological laboratory, experiments were run more in the spirit of today's economics labs rather than psychology labs. An explicit experimental script went without saying – because the experimental subject was Professor Wundt himself or someone else who held a Ph.D. (Danziger 1990). The idea of routinely studying undergraduates rather than experts would have been seen as science fiction, and not very good science fiction at that. The experimenter was merely a technician who controlled the instruments, whereas the subject often published the paper. Repeated trials with the same subject were the rule; they allowed the observation of intra-individual error and systematic changes in performance. Performance-contingent payment was not necessary in order to enhance attention and achievement; the choice of experts as subjects guaranteed sufficient intrinsic motivation. Finally, deception was impossible since the experts knew the script and understood the purpose of the experiment.

In B. F. Skinner's laboratory three-quarters of a century later, a script in the literal sense of a written instruction was not applicable, but trials were repeated and conditions well-controlled enough that even a pigeon could eventually figure the script out. Performance-contingent payment was the rule and, moreover, a central concept of Skinner's theory of reinforcement schedules. Deception in the sense of misinforming the pigeon about the purpose of the experiment was hardly possible.

Wundt's and Skinner's research programs can scarcely be more different in nature. Nevertheless, they illustrate that there have always been practices akin to the four guidelines of today's experimental economists. Therefore, H&O's call for rethinking experimental practice should not be simply put aside by "disciplinary" arguments, such as: OK, economists do different things that demand different practices; that's not psychology, so let's return to business as usual.

A bear market for control. In some parts of cognitive and social psychology, we seem to live in a bear market for experimental control. The reasons for this devaluation can be traced beyond the four practices described by H&O. For example, consider the following puzzling observation. In studies of reasoning, German-speaking students have often been reported as performing at higher levels and more consistently than American students do. For instance, the proportion of Bayesian answers elicited with natural frequencies was substantially higher with German-speaking students (Gigerenzer & Hoffrage 1995) compared to American students (e.g., see Gigerenzer & Hoffrage 1999, Fig. 3). The proportion of students showing perspective change in the Wason selection task was higher with German students (Gigerenzer & Hug 1992) than in most follow-up studies. The same holds for the proportion of students who reasoned according to the conjunction rule in the Linda problem when the question was phrased in frequencies (Hertwig & Gigerenzer 1999). Note that Americans and Germans received the same reasoning problems. What is the explanation for this difference?

Prominent American colleagues have suggested that the cross-Atlantic gap in performance could be due to the higher intelligence of German students compared to American students. Others have suggested that the reason might be the higher average age of German students.

I propose an explanation that attributes the puzzling performance difference to experimental practice rather to the students' traits. In our lab, we typically run participants one by one, or in small groups. Engaging in face-to-face (or "monitor-to-face") con-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

tact with each participant, the experimenter can make practically sure that each participant understands the task or script and that the participant is not distracted and can focus her attention on the task at hand. In contrast, experimenters who reported substantially lower performance generally did not study participants individually. Their students were tested in large classrooms or even in "take-home experiments." The take-home experiment is a recent development in the art of fast data collection. Here, the researcher distributes a booklet with reasoning tasks in the classroom, asks the students to take it home, try to solve the tasks at home, and return the solutions later. Testing students in large classrooms necessarily means losing experimental control, and take-home tests probably mean losing even more. A researcher has no way of knowing under what conditions the student attempted to solve the tasks at home – some students may have been faithful to the instructions, others may have tried to be, but were distracted by noise or interrupted by friends.

My hypothesis is that this loss of experimental control causes, in part, the differences between the performances of German students and American students. This hypothesis can be experimentally tested by systematically studying the effect of one-by-one testing, large classroom studies, and take-home experiments, while keeping culture constant. I would be curious to learn what H&O think of and know about take-home and large classroom studies as a potential factor number 5 in their list of anti-control devices.

Note that the problem of control has already gained a new dimension: data collection on the internet. The internet offers a rapid way to collect large amounts of data. Little, however, seems to be known about the circumstances under which the participants respond on the net, and how these affect the reliability of the resulting data.

Why laissez-faire here and control there? Let me end with another puzzling fact. Cognitive and social psychologists practice laissez-faire, as described by H&O, but at the same time care a great deal about enforcing strict rules for other parts of experimental methodology. For instance, psychologists tend to insist upon the randomized control group experiment as the only legitimate form of experimentation and null hypothesis testing as a "must" for statistical analysis. However, Fisher's randomized group design is only one of several experimental practices used today in the sciences. For instance, another is the demonstration experiment, in which one makes something happen – without the statistical principles of randomization and repetition. This type of experimentation is known from Gestalt psychology, such as when an experimenter tinkers with the spatial and temporal relations between two points of light to produce the phi-phenomenon; it is as prominent in Newton's *Opticks* as in today's molecular biology. Similarly, Fisher's null hypothesis testing is only one form of statistical analysis, and a poor one.

Thus, even within cognitive and social psychology, laissez-faire attitudes and a strict enforcement of rules go hand in hand. The big question is, why laissez-faire here and strict control there? Part of the story seems to be historical accident, followed by blind loyalty to institutionalized habits. Or is there a hidden logic?

A good experiment of choice behavior is a good caricature of a real situation

Francisco J. Gil-White

Solomon Asch Center for the Study of Ethnopolitical Conflict, University of Pennsylvania, Philadelphia PA 19104. **fjgil@psych.upenn.edu**

Abstract: I argue that (1) the accusation that psychological methods are too diverse conflates "reliability" with "validity"; (2) one must not choose methods by the results they produce – what matters is whether a method acceptably models the real-world situation one is trying to understand; (3) one must also distinguish methodological failings from differences that arise from the pursuit of different theoretical questions.

I speak as a psychological anthropologist who uses both psychological and economic experimental methods (lab and field), but who is more familiar with the psychological literature. In general I liked the paper, but I make the following criticisms.

Hertwig and Ortmann (H&O) accuse experimental standards in psychology of being too "diverse." They claim that the "wider range of practices" which they see as coextensive with a "lack of procedural regularity and the imprecisely specified social situation 'experiment' that results may help to explain why 'in the muddy vineyards' (Rosenthal 1990, p. 775) of soft psychology, empirical results 'seem ephemeral and unreplicable' (p. 775)."

Diversity of methods is orthogonal to the precision with which one specifies the social situation "experiment." In principle, one can have an infinite variety of methods, all of which carefully specify it, but in different ways. Likewise, one may have a narrow set of experimental procedures every last one of which fails to specify adequately the social situation "experiment." The criticism that psychologists often fail to specify this situation properly is sound, but this must not be confused with the issue of method diversity, which is a strength of the sociological traditions in psychology.

I see here a conflation of the concepts of *reliability and establishment of validity,* and the impression is strengthened by the exclusive reference (preceding quote) to replicability. In one of the most cited papers in all of psychology, Campbell and Fiske (1959) made the distinction very clearly. In the limit contrast, "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods [replication]. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods" (Campbell & Fiske 1959). When replicability is high, we learn that our methods are reliable. But we know our constructs are good only when validity is high. In general, independent lines of converging evidence are the only way to establish increasing validity for a given claim, and for this we need a variety of methods which will – independently of each other – test it. In one of the best recent examples, Nisbett and Cohen's (1996) claim that the American South has a stronger "culture of honor" than other regions receives empirical confirmation through a variety of mid-level hypotheses, each tested with different methods. The claim thus achieves very high validity. There is no such thing as "too many methods" – on the contrary, just good and bad. And high replicability with low method diversity teaches us about our methods, not about the world.

The argument that because payments to subjects and the use of scripts significantly affect performance, they should be the norm in decision experiments stands on a dubious principle. If a good theoretical model is a good caricature of a real causal process, then a good experiment is a good caricature of a real situation, and *this* should be the standard for the desirability of methodological norms – not whether payments to subjects bring results closer to normative economic predictions, say. The dependent variable is up for investigation, one hopes, and so we can't have methods be chosen according to what kind of quantitative or qualitative results they produce. If payments and scripts affect performance, then at the very least they are something to control for. The case for payments and scripts as methodological norms should stand or fall exclusively on whether they make experiments more like the real world situations we try to understand (this case is easily made). Consider that if increasing similarity to real world situations does *not* affect performance, this is still interesting, still data about how the mind works, and still something to explain. And this implies that the judgment of "reality" should be independent of the measurement of performance in the experiment. Again: the dependent variable is the dependent variable.

H&O argue for the norm that gives participants multi-round experience in a game as if it were the logical solution to the problem of understanding the strategic aspects of the game. But these are

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

separate issues, and the wisdom of multi-round games depends on the research question. If you are interested in people's guesses of opponent performance then you don't want to give them evidence of how others *do* play because this forsakes the insight you were after – people's guesses. Using hypothetical role-playing that exemplifies the strategic possibilities can teach game strategy to participants without experience. I used this method myself in a field experiment (Gil-White 2001). Playing the game several times confounds the issue of learning the strategic structure of the game with learning how to play given the strategies other people actually and typically follow. If – controlling for an understanding of strategy – we get different results with multi-rounds than without them, this may be because the experiment is not a good caricature of life and that, consequently, people's initial intuitions about opponent behavior are initially off-base (although they may adapt to it). Why were their intuitions wrong? One must not confuse a test of learning ability in the experimental situation with a test of how evolved psychology – shaped and modified by the real games we play in everyday life – equips people with intuitions about behavior.

The preceding point applies to feedback in tests of Bayesian reasoning. If (1) people are given rewards for being good Bayesians, and (2) they receive feedback that is immediate and highly accurate, should we – upon the observation of results consistent with Bayesian reasoning – conclude that we have shown that people are good Bayesians, or that experiments set up in this way can train them to be? Again, we must distinguish between people's biases for navigating the world (e.g., their ability to reason abstractly about probabilities) and their ability to learn certain tasks. Feedback is not the solution to a methodological failing, it is merely the tool required for a particular kind of question. The next question is: How good a caricature of real-life feedback processes are the feedback mechanisms used in these experiments?

Theory-testing experiments in the economics laboratory

Anthony S. Gillies[a] and Mary Rigdon[b]

[a]*Department of Philosophy and Program in Cognitive Science;* [b]*Department of Economics and Economic Science Laboratory, The University of Arizona, Tucson, AZ 85721.* **agillies@u.arizona.edu**
www.u.arizona.edu/~agillies **rigdon@econlab.arizona.edu**
www.u.arizona.edu/~mlrigdon

Abstract: Features of experimental design impose auxiliary hypotheses on experimenters. Hertwig & Ortmann rightly argue that the ways some variables are implemented in psychology cloud results, whereas the different implementations in economics provide for more robust results. However, not all design variables support this general conclusion. The repetition of trials may confuse results depending on what theory is being tested. We explore this in the case of simple bargaining games.

The conclusions that we draw from theory-testing experiments are tempered by what auxiliary hypotheses we have about the problem domain that are relevant to the experiment. If a theory predicts that *P,* and in our experiment we observe *not-P,* before we conclude that the theory does not square with the facts we must be able to rule out that it is some auxiliary hypothesis *H* (perhaps a hypothesis about how our experiment is run) tacitly appealed to that is confounding the prediction of the theory (Duhem 1906; Quine 1953b). Hertwig & Ortmann (H&O) rightly draw attention to a special case of this fact: the difference between experiments in economics and in similar branches of psychology. While we think the target article gets the story mostly right, it does misclassify how one design variable in particular (repetition of trials) interacts with different theories. The use of deception, financial incentives, and scripting all seem to be general variables whose appropriateness is independent of the theory being tested; repetition of trials may confuse rather than clarify experimental results depending on the target theory.

Experimental practices in economics and psychology differ and, in particular, some of the differences can be described as varying the extent to which procedures serve as confounding hypotheses in theory-testing. If we are interested in rational choice (both in pure decision problems and strategic environments), why should we think that what subjects say they would do should track what they actually would do if given financial incentives? After all, that is what makes choices interesting in real life: something turns on what it is we choose. The lack of financial incentives clouds experimental results; an experiment in which subjects are paid a flat fee imposes a rather implausible auxiliary hypothesis into the story. Similar problems arise for some of the other design variables that H&O discuss, like the use of deception by experimenters and providing a well-defined decision environment (what they call a "script"). Deceiving subjects, for instance, imposes the implausible auxiliary hypotheses that experimenters have no reputation among subjects and that subjects will complete a given task in an experiment even if they are suspicious about what the "real" task of the experiment is. The basic conclusion for experiments without a clearly defined script is roughly the same.

The same general sort of conclusion cannot be drawn with respect to whether an experiment allows for repeated trials or not, as H&O mistakenly think. We are most interested in the case of experimental games. The authors argue that "equilibrium is assumed not to be reached right away" but that "it is expected to evolve" over time until a steady-state in observed behavior is reached. And this is meant to be the reason why economists insist on the use of repeated trials. This is erroneous for at least two reasons.

First, this argument gets the history of game theory backwards. The traditional concept of equilibria is that they represent where ideally rational agents would play, given certain information conditions. These agents, being ideally rational, can be expected to reach an optimal outcome immediately. Equilibria, in this classical interpretation, are supposed to characterize the set of possible outcomes that such agents could reach by deliberating about the structure of the game. On the other hand, the "learning" interpretation of equilibria in games is quite recent, and represents "the alternative explanation that equilibrium arises as the long-run outcome of a process in which less than fully rational players grope for optimality over time" (Fudenberg & Levine 1998, p. 1). So it is a mistake to think that the traditional interest in equilibria *requires* that experiments use repeated trials.

Second, depending on what theory a given experiment is meant to be testing, repeated trials may impose rather than eliminate an implausible auxiliary hypothesis into the story. Take the case of simple two-person trust games (Berg et al. 1995). H&O suggest that the "attention-grabbing results of these games [might be] due to the very fact that they are typically implemented as one-shot rather than repeated games." They offer a distinction between two sorts of repeated environments, stationary replications and what may be called repetition with replacement. But if we take the suggestion that our experimental design imposes auxiliary hypotheses into the mix seriously, we have to consider whether either of these sorts of repeated trials is appropriate to the theory being tested. In the case of bargaining experiments, where the understood theory being tested is the standard noncooperative solution concept of Nash equilibrium and its refinements, neither of these types of repeated trials is essential.

Consider first, repetition with replacement. In these games, subjects may meet one another more than once. The future may loom large enough that cooperative play is supported by the Folk Theorem. But then, observing cooperative behavior in this environment would not distinguish between noncooperative game theory and more exotic solution concepts (e.g., trust and reciprocity). So if we want to test the noncooperative story, this environment is not best suited to that purpose.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Now consider stationary replications. This sort of environment allows for subjects to play multiple periods, but with different partners. Again, however, it is far from obvious that this type of design is the most appropriate for theory-testing in bargaining environments like a simple two-person trust game. First, using stationary replications in a trust game imposes the rather implausible auxiliary hypothesis that subjects never update their estimates of the distribution of player types (trusters or not) in the environment. Without this hypothesis, the experiment would not provide a direct test of the noncooperative theory since there is a noncooperative equilibrium that is predicted even with great uncertainty about the other players in the population. Second, the empirical facts suggest that cooperation can be sustained with stationary replications in trust games (McCabe et al. 1998; 2000).

The point we want to make is that, for theory-testing experiments, *what* the theory is dictates whether repeated trials are appropriate or not. For a variety of experimental games, one-shot experiments give noncooperative theory its best shot at predicting behavior. This is not to say that repeated trials are either uninteresting in these games, or never appropriate for theory-testing. Repetition of trials, we suggest, is just a different sort of design variable from deception, scripting, and the use of financial incentives.

The contribution of game theory to experimental design in the behavioral sciences

Herbert Gintis

Department of Economics, University of Massachusetts, Amherst, MA 01003.
hgintis@mediaone.net **www-unix.oit.umass.edu/~gintis**

Abstract: Methodological practices differ between economics and psychology because economists use *game theory* as the basis for the design and interpretation of experiments, while psychologists do not. This methodological choice explains the "four key variables" stressed by Hertwig and Ortmann. Game theory is currently the most rigorous basis for modeling strategic choice.

"Why do the methodological practices differ?" ask Hertwig and Ortmann (H&O). They answer: "There is no simple answer to this question." But there is a simple answer: economists use *game theory* to design and interpret experiments. This choice explains virtually every aspect of experimental economics, including how the "four key variables" stressed by H&O tend to be implemented.

Game theory is not a "theory" at all, but a general framework for modeling strategic action and interaction. Game theory does not predict behavior. Rather, it models the characteristics of players, the rules according to which they interact, the informational structure available to the agents, and the payoffs associated with particular strategic choices. As such, game theory fosters a *unified approach to behavioral science*, allowing researchers to employ the same language and techniques, whether their training be in biology, anthropology, psychology, or economics. Game theory promotes well-controlled experimental protocols, so experimental conditions can be systematically varied and experimental results can be replicated in different laboratories (Plott 1979; Sally 1995; Smith 1982).

Many researchers have rejected game theory because game theory's rational self-interested agent does not capture most human behavior. However, experimental findings in recent years have led economists to broaden their model of the human actor considerably. Indeed, contemporary game theory treats agents' objectives as *facts to be discovered* rather than *behavior deduced from the laws of reason.* Game theorists have accordingly built and tested models of regret, altruism, vindictiveness, status-seeking, trust, and other broad individual and social behaviors that do not fit the model of hyper-rationality (Gintis 2000, Ch. 11).

Consider the four key variables stressed by H&O.

Enactment of scripts. Game theory suggests that agents optimize subject to constraints, given the information they possess concerning payoffs and the nature of the other players. Therefore, behavior will be sensitive to the manner of presentation of the experimental situation to the subjects. The ability to control experimental conditions and replicate experimental results therefore depends on scrupulous attention to the wording of instructions given to subjects, as well as the realization of the physical and social environment of the experiment.

For example, consider the Ultimatum Game (Güth et al. 1982). A subject, the Proposer, is told to offer between one and ten dollars to another (anonymous and unseen) subject, the Respondent. If the Respondent accepts the offer, the money is divided up accordingly. If the Respondent rejects the offer, both Proposer and Respondent receive nothing. If both parties are rational and self-interested, game theory predicts that this amount of information, given to both subjects, is sufficient to predict that the Proposer will offer one dollar, and the Respondent will accept. Yet when this experiment is actually run, most Proposers offer considerably more that one dollar (the average offer is often around four dollars), and low offers (e.g., of less than three dollars) are likely to be rejected. Respondents appear to reject offers they feel are unfairly low, and Proposers, expecting this behavior, make offers unlikely to be rejected. Because Respondents care about "fairness," the wording of the instructions affects subject responses. For instance, even if both Proposer and Respondent are given the same instructions, their behavior may be erratic unless each *knows* that the other was given the same instructions. Or, for instance, if the Respondent is not told how much money the Proposer has to divide between the two (e.g., the Proposer could have as little as one dollar or as much as one hundred dollars), then experiments show that the Respondent will accept much lower offers than with complete knowledge. The Proposer must thus be given the information that the Respondent knows that the total is exactly ten dollars.

More generally, even when there are clear monetary payoffs, if the experimental conditions permit subjects to reward and punish one another, most subjects do not maximize their monetary return. In such cases subjects are strongly influenced by the way the experimental situation is "framed." For instance, subjects may be called "opponents" or "partners," they may be told to "contribute" or to "allocate" funds, and they may be allowed to "punish" or to "assign points" to other subjects. Different wording produces different behavior. Behavior is sensitive to instructions not only because they affect subjects' objective functions, but also because each subject knows that the instructions will affect the behavior of the other subjects, and adjust personal behavior accordingly.

Repetition of trials. Successful experimental design must carefully plan the type and number of repetitions of encounters. Game theory suggests two reasons. First, since agents optimize subject to constraints and to the behavior of others, agents' initial responses to an unfamiliar situation are unlikely to represent their responses after a learning period. Second, the nature of repetition affects the optimal strategies themselves. One well-known example of this is the Prisoner's dilemma, in which mutual defection is optimal when the game is repeated a fixed number of times, but in which cooperation is optimal if the game is repeated with a positive probability of continuance, provided the agent's discount rate is sufficiently low.

Performance-based payments. From a game-theoretic perspective, a careful specification of payoffs is a prerequisite for predicting agent behavior. From this derives the stress on explicit rewards and penalties in experimental design. It is important, however, not to presume that the experimenter's explicit payoffs alone enter the subject's objective function. For instance, as we have seen in the Ultimatum Game, even with explicitly monetary payoffs, fairness and reciprocity enter the subjects' objective functions. Perhaps the most important reason to have monetary or other explicit payoffs is to avoid a situation is which the subject, who, for reasons described by H&O, places excessive emphasis on satisfying what the subject considers to be the wishes of the ex-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

perimenter. Monetary incentives do not completely "crowd out" the subject's motive to please, but they are likely to attenuate this motive considerably.

The avoidance of deception. According to game theory, in a novel situation agents attempt to determine what sort of game they are playing, and adjust their behavior accordingly. Experimenter deception adds a degree of uncertainty to this assessment, and hence adds a dimension of uncertainty to their behavior and to the interpretation of the results of the experiment. Therefore, experiments are more likely to achieve interpretable results if the experimenter can render it extremely probable in the subjects' minds that this dimension of strategic action is absent from the experiment.

ACKNOWLEDGMENT

I would like to thank the MacArthur Foundation for financial support.

Are scripts or deception necessary when repeated trials are used? On the social context of psychological experiments

Adam S. Goodie

Department of Psychology, University of Georgia, Athens, GA 30602-3013.
goodie@egon.psy.uga.edu
www.teach.psy.uga.edu/Dept/Faculty/Goodie/goodie.stm

Abstract: Scripts and deception are alternative means, both imperfect, to the goal of simulating an environment that cannot be created readily. Under scripts, participants pretend they are in that environment, while deception convinces participants they are in that environment although they are not. With repeated trials, they ought to be unnecessary. But they are not, which poses challenges to behavioral sciences.

Hertwig and Ortmann (H&O) outline a number of valid generalities about the relative use of four key variables in the disciplines of psychology and economics that, due to their breadth, are bound to have significant exceptions, as the authors acknowledge. At the same disciplinary level of generality, and with the same acknowledgment that exceptions will exist, three of these variables may be drawn together thus: With repeated trials, scripts ought to be unnecessary and deception ought to be pointless. But they are not so, which poses challenges to behavioral sciences in general.

Scripts and deception may be viewed as alternative means to the common goal of simulating response in an environment that cannot be created readily. Scripts ask participants to pretend they are in that environment, while deception aims to convince participants that they really are in that environment even though they are not. Both have pitfalls. H&O present clear evidence that deception often fails to convince participants of its pretense – and even makes it difficult for non-deceiving experimenters to convince participants of their honest settings.

The use of scripts – enactment, role-playing and the like – also has dangers. It is at its core a theatrical enterprise, as is suggested by the terms used to describe it. If an experimenter asked participants to act as if a loved one had died, we would not assume that the resulting behavior provided a faithful picture of the grief process. Is it safe to assume that a participant trying to behave as if she were a stock trader provides a true picture of trading? On the face of it, it is not. H&O provide some reason to think that role-playing yields representative results, namely, that behavior is brought closer to normative standards and is made less variable. Still, in any particular experiment, one does not know if scripted interactions are representative of non-scripted ones, and it is generally hazardous to use one's primary data as a manipulation check.

The use of repeated trials should greatly reduce the need for either deception or scripts. It is not necessary to ask participants to imagine they are engaged in a particular task, or to convince them they are engaged in a particular task, when they actually *are* engaged in the task. Why must a participant try to be like a bond trader in pursuit of maximal money, when a participant really is a student playing a computer game for maximal money in the context of a psychological experiment? One can simply provide participants with enough information to begin playing the game, and study the strategies that emerge as the game is played (Goodie & Fantino 1995; 1996; 1999a; 1999b). Experiments are not outside of participants' lives; they are a small part of them. The fact that behavior in experiments fits into patterns of behavior in participants' lives can lead experimenters into error – for example, speciously detecting short-run bias in what is part of an effective long-run strategy (Arkes & Ayton 1999; Dunwoody 2000; Goodie & Fantino 1999a) – but it can also be an asset that provides well-learned initial conditions (Goodie & Fantino 1996).

Interesting to note, a script element that H&O emphasize is the explicit instruction to maximize performance, whether performance is rewarded with money, points or mere correctness. They document that such instructions often change performance, which suggests that the real incentive for experimental participants is not money, points or correctness, but compliance with instructions instead. When instructed to maximize profit, participants maximize profit (or approximately so); but when instructed otherwise, they do otherwise. This is relevant to Roth's (1995) framing of the difference between experiments that use financial incentives and those that do not as "the question of actual versus hypothetical choices" (p. 86), which H&O adopt. If money must be backed by compliance in order to be an effective incentive, as is sometimes the case, then some caution is warranted in deciding what incentives to use, and the question is not as simple as "financial incentives versus no incentives." In some settings, a financial incentive might be no incentive, and other incentives might be real incentives. More generally, choices motivated by points, tokens, the approval of the experimenter or the (learned or unlearned) rewards of correct answers per se are actual choices, not hypothetical ones. Such choices may or may not be as strongly motivated as those that earn money, but they are not less real.

It complicates matters considerably that although scripts should not matter when repeated trials are used, they often do matter. Performance on tasks that are instructed is less adaptable to changes in the rules than performance on tasks that are learned by experience (Lowe 1980; Matthews et al. 1977). And H&O document a number of findings to demonstrate that performance is sensitive to instructions, even under repeated trials. Choices must be made between studying the effects of rules of the game and the effects of instructions, and also about assumptions that will be made about the variable or variables not under investigation. One cannot study the effects of the rules of the game, which are typically identified as completely determining normative standards of performance, without having one's data affected by the scripts chosen. H&O have done the field a significant service by highlighting the often neglected and often complex but unavoidable impact of scripts on performance.

Clear-cut designs versus the uniformity of experimental practice

Francesco Guala

Centre for Philosophy of the Social Sciences, University of Exeter, Exeter EX4 4RJ, United Kingdom. **f.guala@ex.ac.uk**
www.exeter.ac.uk/shipss/sociology/staff/francesco/index.html

Abstract: Clear-cut designs have a number of methodological virtues, with respect to internal and external validity, which I illustrate by means of informal causal analysis. In contrast, a more uniform experimental practice across disciplines may not lead to progress if causal relations in the human sciences are highly dependent on the details of the context.

Hertwig and Ortmann's paper features two messages that should be kept distinct. The first one is a plea for more uniform experi-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

mental practice across neighbour disciplines like economics and psychology; the second one is a plea for more clear-cut designs in these areas. It emerges from the last part of the paper that both proposals are aimed at a unique – and indeed extremely important – goal, namely, enhancing the replicability and comparability of results across fields, and that they eventually promote progress in the "softer" discipline of psychology. Hertwig and Ortmann (H&O) also seem to suggest that their proposals be hierarchically ordered, the plea for clear-cut designs being functional to the achievement of the other goal (uniformity), which would in turn be instrumental to the achievement of the higher-order desideratum (scientific progress).

I would like to argue that we should revise the order of the two proposals: clear-cut designs must definitely come first, uniformity of experimental practice being much more questionable as a methodological standard. By "clear-cut design" I mean an experimental set-up that maximises control. At a very general level of analysis, this can be achieved in two ways: (1) by keeping other factors in the background of the independent variable stable, and/or (2) by randomisation. Letting other factors vary in an uncontrolled fashion raises the worry that we get caught in "Simpson's paradox" situations. But even if we were able to identify a genuinely causal, non-spurious relationship, letting too many things vary in the background might confound the result by unduly inflating the error term and hiding the significance of the main effect.

All these worries have to do with internal validity. But a clear-cut design also helps to deal with external validity in a most effective way. Suppose an experiment investigates the proposition that factor X is the (or a) cause of Y. In order to achieve internal validity, we shall check whether exogenous manipulations of X are reflected in variations of Y, controlling for other experimental parameters – for example, the kind of setting, of subjects, and so on (W, Z, . . .). A clear-cut design will test the effect of X on Y, keeping W (or Z, etc.) fixed at one level (often at the zero level); the same procedure can be repeated at another level, and so on, until we have exhausted the plausible sources of variation. Notice that letting the background vary (or randomising across different background factors) might not *prevent* us from measuring the effect we are interested in. In particular, variations will not do too much harm if the effect of X combines additively with the effects of the other factors on Y; X's net contribution might be partly confounded by the other factors, but will nevertheless be constant across the experiment(s). An additivity assumption solves so many problems (including problems of statistical analysis) that it is often taken for granted in empirical research and even in philosophy of science. But consider the possibility of interactions: X "interacts" with Z if its contribution to Y differs depending on the level of Z. If interactive effects are common, two consequences follow: first, randomisation (with background variations) is likely to result in misleading results, in that the measurements may not reflect the effect of X on Y in any particular causal set-up. Secondly, clear-cut designs will in contrast provide genuine causal information about the effect of X on Y in specific causal contexts (say, when Z is set at level Z*, W at W*, etc.) but will not be generalisable outside that more or less narrow domain.

Thus, the relative advantage of clear-cut designs can be rationalised using considerations of causal inference (more on this in Guala 1999; 2000). Following the above reasoning, experimenting with *and* without repetition, with *and* without monetary incentives, and with *different* clear-cut scripts, should provide us with more reliable causal information independently of the ontology of the subject matter. (The ban on deception belongs to an altogether different set of considerations and I will not comment on it here.) But in H&O's paper the "do-it-both-ways" rule is intended to promote uniformity in experimental practice and, eventually, theoretical convergence. The idea is to create a data-base of results for various categories of experiments, providing rules like "if you do an experiment of type A, then remember that financial incentives matter" (and, perhaps, you won't be published if you don't implement them); "if you do an experiment of type B, remember that repetition is important"; "for experiments of type C, in contrast, we have had mixed results, so do it both ways," and so on.

Such an institution, or something like it, would probably be successful in promoting more uniformity and theoretical convergence in areas that presently lack it. But what we want is not convergence per se. We seek convergence towards true causal knowledge, and the truth may be that causal relations are more context-dependent than we would like to think. Indeed, a lot of experimentation in cognitive psychology and behavioural decision-making is prompted by the idea that behaviour is context-dependent, and sensitive to more differences 'in the background' than suggested by standard economic theory. One can easily foresee some disagreement, when it will come to define the domain of "type A," "type B," and such other experiments with their respective methodological canons.

We now know that financial incentives have unexpected effects on the rate of preference reversals (Grether & Plott 1979) and we would have never discovered it, had we not done it both ways. But we may have been prevented from discovering such effects, had we relied on some "similar" experiment suggesting that incentives always improve cognitive performance. Likewise, focusing on repeated trials may prevent us from investigating a whole domain of behaviour that is interesting on its own (and the excuse that economics only deals with "equilibrium" behaviour is quite poor: economic theory should be, and is in fact used to provide explanations of economic behaviour in general).

To sum up: we should go for more clear-cut designs, but seek uniformity only if the subject matter allows it. But that depends on how the (social) world is, and I don't see any reason to legislate about it at this stage. Thomas Kuhn (1962) famously argued that a fair amount of dogmatism, tight rules, and conventions are necessary prerequisites for scientific progress. Yet, one should be wary of imposing the paradigm of contemporary economics on other neighbouring areas that have constantly provided economists with challenges in the form of robust anomalous phenomena. As long as that was the result of etherogeneous theorising, and different methodological approaches, it will be desirable not to impose too much uniformity in the next future as well. More *clear-cut and varied* designs, then, is perhaps the best recipe for experimental social science.

Doing it both ways – experimental practice and heuristic context

Glenn W. Harrison and E. Elisabet Rutström

Department of Economics, The Darla Moore School of Business, University of South Carolina, Columbia, SC 29212.
{harrison; lisar}@darla.badm.sc.edu
www.dmsweb.badm.sc.edu/{glenn; lisa}

Abstract: Psychologists can learn from the procedural conventions of experimental economics. But the rationale for those conventions must be examined and understood lest they become constraints. Field referents and the choice of heuristic, matter for behavior. This theme unites the fields of experimental psychology and experimental economics by the simple fact that the object of study in both cases is the same.

We generally agree with the themes proposed by Hertwig and Ortmann (H&O). Control is the hallmark of good experimental practice, whether it be undertaken by economists or psychologists. But practices meant to control may also blunt some of the very behavior that one would like to study. So while we endorse the prescriptions to psychologists that they employ some standards that allow results from one experiment to be replicated and modified incrementally, we would caution against too dogmatic a line. We take their main advice to be summarized in the adage to "do it both ways," by which we understand them to be encourag-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ing experimenters to replicate the previous design and then study the effects of orthogonal variations.

Unfortunately, experimental economists have sometimes followed conventional practices with little thought about the consequences.

Consider, for example, the popular use of "lab dollars" (Davis & Holt 1993, pp. 225ff.). These are a lab currency used in the experiment itself, and then converted to some local currency at the end of the experiment. Invariably, these lab dollars have lots of zeroes after them, so that instead of bidding $30 one might observe a subject bidding 30,000 "lab pesos." The purported reason for using this device is to give the subjects greater incentive to report monetary responses at a finer level of detail than if a field currency were used. The problem is that this will occur only if the subject suffers from some illusion with respect to the exchange rate between lab currency and field currency. Because such illusion is bound to vary across subjects, one has lost control over incentives. At the very least, the incentives will be much lower than intended, reducing saliency and increasing noise in the data. In the worst case, payoff dominance problems may cause results to be biased (Harrison 1989).

H&O consider the need for standards in four areas: script, repetition, financial incentives, and deception. We comment on the first two.

One tradition in experimental economics is to use scripts that abstract from any field counterpart of the task (Camerer 1995, pp. 652ff.). The reasoning seems to be that this might contaminate behavior, and that any observed behavior could not be then used to test general theories. There is a logic here, but we believe that it may have gone too far. Field referents can often help subjects overcome confusion about the task. Confusion may be present even in settings that experimenters think are logically or strategically transparent. If the subject does not understand what the task is about, in the sense of knowing what actions are feasible and what the consequences of different actions might be, then control has been lost at a basic level. In cases where the subject understands all the relevant aspects of the abstract game, problems may arise due to the triggering of different methods for solving the decision problem. The use of field referents could trigger the use of specific heuristics from the field to solve the specific problem in the lab, which otherwise may have been solved less efficiently from first principles (Gigerenzer et al. 2000). For either of these reasons: a lack of understanding of the task or a failure to apply a relevant field heuristic, behavior may differ between the lab and the field. The implication for experimental design is to just "do it both ways." Experimental economists should be willing to consider the effect in their experiments of scripts that are less abstract, but in controlled comparisons with scripts that are abstract in the traditional sense. Nevertheless, it must also be recognized that inappropriate choice of field referents may trigger uncontrolled psychological motivations. Ultimately, the choice between an abstract script and one with field referents must be guided by the research question.

Another tradition in experimental economics is to use repetition, typically with some shuffling of subjects from period to period so that specific opponents in each round are not the same person. The logic is that subjects need to become familiar with the task before we can say that their behavior is a reliable measure of their responses to that task (e.g., Holt 1995, pp. 403ff). The citation from Binmore (1994, pp. 184ff) is right on the mark: Repetition is a device developed by experimental economists to provide some operational counterpart to game theorists' notion of common knowledge. But so stated, it is just one possible device, and not always the best one. In addition to the traditional statistical problem of serial correlation, traditionally ignored by experimental economists, it is also possible that repetition encourages subjects to use different heuristics than they would use if just playing the game on a one-shot basis. For example, if subjects are told that they will play some game for 100 trials, that each trial will last no more than 20 seconds, that there will be some feedback on performance, and that total payoffs will be the simple sum of payoffs over all trials, this might encourage use of a heuristic that says "start with any old solution and then iterate by means of local perturbations toward one that gives me better feedback; stop perturbing the solution value when the increment in reward drops below some threshold." This heuristic could easily miss some global optimum when the payoff function is not extremely well-behaved (Harrison 1992). If, instead, the subject is told that she will have only one shot at the game, and all the payoff comes from this one solution, a different heuristic might be triggered: "exhaustive evaluation of all (discretized) choices." This heuristic would be much more likely to identify a global optimum.

Psychologists obviously have a lot to learn from the new experimental kids on the block. But we do not believe that the intellectual trade is all one way. Field referents and the choice of heuristics matter for behavior. This is a simple and magnificent theme that unites the fields of experimental psychology and experimental economics, by the sheer fact that the object of study in both cases is the same. The gentle aggression of H&O moves us much closer to a common language for scientific discovery.

ACKNOWLEDGMENT

Rutström thanks the U.S. National Science Foundation for research support under grants NSF/IIS 9817518, NSF/MRI 9871019, and NSF/POWRE 9973669.

Challenges for everyone: Real people, deception, one-shot games, social learning, and computers

Joseph Henrich

Business School, University of Michigan, Ann Arbor, MI 48109-1234.
henrich@umich.edu **www.webuser.bus.umich.edu/henrich**

Abstract: This commentary suggests: (1) experimentalists must expand their subject pools beyond university students; (2) the pollution created by deception would not be a problem if experimentalists fully used non-student subjects; (3) one-shot games remain important and repeated games should not ignore social learning; (4) economists need to take better control of context; and (5) using computers in experiments creates potential problems.

Hertwig and Ortmann (H&O) provide a useful synthesis and comparison of experimental methods from economics and psychology (i.e., a particular sub-field of psychology). However, in reflecting on methodological improvements, H&O fail to recognize the extreme reliance of both psychology and economics on university students as subjects. Many scholars from both fields are guilty of reaching conclusions about "human reasoning" (i.e., conclusions about our species) by relying entirely on this very weird, and very small, slice of humanity. Even efforts at cross-cultural work by both economists (e.g., Cameron 1999; Roth et al. 1991) and psychologists (Nisbett et al., in press) nearly always involve university students, albeit from places like China, Indonesia, Japan, and Israel. Meanwhile, there are three billion potential adult subjects out there who have never participated in any experiments, and are a lot more representative of humanity than the privileged inhabitants of elite universities. Researchers do not need to go to the Amazon, Papua New Guinea, the Ituri Forest, or Kalimentan – try the bus station, the beaches, the market, the used furniture auction, the bowling alley, the "projects," or the county fair.

H&O also express an important concern about how the repeated use of deception might adversely affect experimental results. However, if researchers would simply expand their potential subject pools to include the three billion adults who have never set foot in a psychology or economics class (or in a university), all worries about polluting the human subject pool could be put to rest. All deception experiments could be done outside the univer-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

sity, with a new set of subjects every time if need be. If every experiment requires 100 fresh subjects, we could do 30 million experiments before having to use anyone twice.

I agree with H&O that repeated games can be useful, but they provide little actual defense for their emphasis on repeated games, aside from quoting Binmore. First, real life is full of one-shot games in which individuals must make one-time decisions without any prior experience. Many people get married only once (divorce remains illegal in many countries). Many families buy or build only one house. People also often get only one chance to deal with a breast cancer diagnosis, a physical assault or a sinking ship. Second, in some experiments (like the Ultimatum Game), repeating the game does not substantially affect the results (Roth et al. 1991). Third, sometimes we only want to measure what people bring into the experiments, not what they can learn via experience in highly structured laboratory settings with clear feedback. I think both one-shot and repeated games are useful, and there is no particular reason to emphasize one or the other until a research question is specified.

The other problem with repeated-game experiments is the almost complete emphasis on studying individual learning, as opposed to social learning. We know that if you provide experimental subjects with opportunities to imitate others, they will (Bandura 1977; Henrich & Gil-White 2001), especially when money is on the line (Baron et al. 1996; Kroll & Levy 1992). In real life, individuals are repeatedly exposed to the behavior and decisions of others; consequently, it is a serious methodological concern that economics experiments usually ignore the potentially important impact of imitation and other forms of social learning on adaptive learning.

H&O emphasize the importance of detailed instructions, scripts and context. I fully agree. As the authors hint, many economists go to extreme lengths to excise any context from their experiments in an effort to make the payoff structure the centerpiece of decision-making. However, if people use contextual cues to figure out which set of norms applies to a particular problem (independent of the game payoffs), then acontextualizing an experimental protocol effectively creates an uncontrolled variable, which liberates subjects to interpret the game according to whatever context comes to mind first (which may be governed by their last social interaction outside the lab, or what they had for lunch). Such acontextualizing may, for example, account for the high variance found both within and across public goods games. The "details" researchers abstract out of their game structures may be exactly those details that cue people into using particular sets of norms. Economists have taken control of a specific set of potentially important variables that relate to payoffs, but they may still be missing other key variables (Pillutla & Chen 1999).

Both psychologists and economists use computers to administer experimental protocols. Unfortunately, psychological evidence indicates that people often think of computers as social actors, with feelings, genders, motivations, and even emotions (Nass et al. 1997). Consequently, experimentalists may be effectively introducing omniscient third parties into their games. Computers may not affect the payoff matrix, but they may affect human psychology.

Is the challenge for psychologists to return to behaviourism?

Denis J. Hilton

Department of Psychology, University of Toulouse-II, 31000 Toulouse, France. **hilton@univ-tlse2.fr**

Abstract: I suggest that contemporary economics shares many of the characteristics of methodological behaviourism in psychology, with its emphasis on the influence of motivation, learning, and situational incentives on behaviour, and minimal interest in the details of the cognitive processes that transform input (information) into output (behaviour). The emphasis on these characteristics has the same strengths and weaknesses in economics as in behaviourist psychology.

In my reactions to Hertwig and Ortmann's (H&O's) article, I begin with some historical perspective. To my mind, H&O's questions are very similar to the ones that behaviourist psychologists would pose about current research in psychology. This is no doubt because economists are still for the most part methodological behaviourists (cf. Lewin 1996), focussing on observables such as market share and prices and eschewing detailed analysis of *real* cognitive processes, preferring to use mathematically elegant but psychologically implausible models to link inputs (information) to outputs (choices). Thus experimental economists, like behaviourist psychologists, emphasise the roles of learning, motivation, and situational forces (reinforcement contingencies, cost-benefit matrices, etc.) in controlling human behaviour.

Although economists have done contemporary psychologists a service by reminding us of the importance of learning, incentives, and environmental forces, neither group should fall into the error of supposing that psychologists have *never* taken the influence of situational factors, learning, and motivation on human behaviour seriously. Today's cognitive psychology is very much a reaction against the shortcomings of the behaviourist approach, and owes its successes to taking the idea that cognitive processes explain important aspects of behaviour and impose structure on the environment, rather than being a *tabula rasa* written on by external forces. Linguistics, anthropology, biology, and computer science adopt similar forms of structural explanation (Piaget 1996), and the fundamental question these life and information sciences pose is to ask how input is transformed into output. Consequently, much experimentation in psychology is directed to understanding these cognitive (transformation) processes.

That economists in recent years have taken so much interest in the results of experimental psychology seems to me to be an encouraging sign for psychologists, and would itself seem to me to argue in favour of psychologists continuing much as before. H&O's valuable article indeed succeeds in raising important questions, but a balanced answer to each one needs to take a wide historical and methodological perspective into account. Even if there is merit in the "quality control" proposed by some experimental economists, it is surely in no one's interest to bridle the creativity that has enabled psychologists to mount such a strong and systematic challenge to the assumption of economic rationality in human behaviour. I therefore deal with the major challenges raised by H&O below, beginning by taking the questions of repetition and incentives together, before going on to discuss scripts and deception.

Repetition and incentives. For me the most important question raised by H&O is whether psychological biases can be eliminated by learning and experience. The evidence strongly suggests that they cannot. Experimental incentives do sometimes have an effect of improving rationality, but not always (only in 23 out of the 43 studies created by combining the Hertwig-Ortmann and Camerer-Hogarth reviews). Even in these studies, irrationality was probably not altogether eliminated (though unfortunately we are not given precise information on this point). In seven others there was no effect of incentives, and in a remarkable 13 cases they had negative effects. Indeed, given that Wall Street bond traders dealing day after day in millions of dollars show irrationalities predicted by prospect theory (Shapira 2000) it would be surprising if small experimental learning and incentives eliminated irrationality. Finally, the frequency of paradoxical effects of reward sounds like a wake-up call to economists to develop richer theories of human motivation, just as Festinger and Carlsmith's (1957) experiments on insufficient justification forced psychologists to address the ways in which cognitive dissonance processes mediate the effects of reward on behavioural intentions.

In conclusion, while the Hertwig-Ortmann and Camerer-Hogarth studies show that cognitive error can be partly corrected by

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

learning and incentives, the larger picture remains true – psychologists did the right thing to abandon behaviourism and to study cognitive processes through one-shot experiments. Doing learning and incentive experiments which take longer to do and are more expensive, do not change the picture radically (and maybe are best farmed out to experimental economists and learning theorists with big research grants). In addition, research needs to focus more on when learning reaches asymptote to assess the stable, lasting effect (if any) of learning and incentives on irrationality.

Scripts. A first issue is that I think that the general point that experimenters should specify the norms in a situation before evaluating performance is a good one. However, it does not follow that psychologists should always follow the directive route of economists of explaining to people what they *should* do. It can also be interesting to study how people might spontaneously structure the situation, for example, by assuming that implicit conversational norms frame the information given, or that fairness norms might be relevant in a game situation. While I agree that the experimenter should assure himself of what norms the participant is using, explicit instruction does not seem to me the only way to do this.

A second issue is that I think that H&O's example of the Wason selection task does not illustrate their point about instruction through scripts well. No matter how hard one tries to explain to people that this is a *logical* task, or to pay them, or try to get them to learn through giving them easier concrete versions to practice with, most will fail on the abstract version. They simply fail to understand the problem in its abstract form, and the fact that "social contract" versions cue them into "correct" interpretations is proof of an automatic *cognitive facilitation* effect, and does not indicate a weakness in the original experimental methodology. The difficulty of the abstract version of the Wason task remains a severe challenge to anyone who believes that people are "inherently logical," or that having a university education guarantees content-independent logical thinking.

Deception. I agree that too much use of deception is a bad thing, but as a teacher of social psychology I also know that the experiments that most challenge students' preconceptions about human behaviour on conformity, obedience, bystander non-intervention, and so on, almost all use deception. While we should minimise the costs of contaminating subject pools through deception, the social benefits of the knowledge gained through the effective use of deception seem to me to clearly justify its use where appropriate.

To what are we trying to generalize?

Robin M. Hogarth

Department of Economics and Business, Pompeu Fabra University, 08005 Barcelona, Spain. **robin.hogarth@econ.upf.es**

Abstract: In conducting experiments, economists take more care than psychologists in specifying characteristics of people and tasks and are clearer about the conditions to which results can be generalized. Psychologists could learn much from this practice. They should also think in terms of programs of research – involving laboratory and field studies – that will allow them to test the generality of their theories.

The paper by Hertwig and Ortmann (H&O) is a welcome contribution to the literature because it forces psychologists – and particularly those of us doing research in judgment and decision making – to think about how we go about our business. Experimental economics is a "boom industry" and it is helpful to compare and contrast its experimental practices with those in psychology.

At a general level, both psychologists and economists share a common theoretical perspective. Behavior is a joint function of the organism and the environment in which it acts. However, the ways in which the two disciplines make this operational differ considerably. In economics, there is a precise model of the organism in terms of utility functions, beliefs that take the form of probabilities, and a rule for action (maximizing expected utility). As to environments, these are typically characterized by costs, payoffs, and specific institutional arrangements. In addition, theory specifies that different structural representations of the environment (e.g., framing of decision problems) should make no difference. In psychology, on the other hand, assumptions about people and environments tend to be vague (although there are notable exceptions), and problem representation is a key issue. Context – however vaguely defined – is important to psychologists, but not to economists.

The practice of experimental economics is exemplary in designing tasks in ways that replicate salient features of the real world situations to which the theories being tested apply. Subjects, for example, are required to behave within well-defined roles and institutional arrangements are meticulously respected. There may be two reasons for this. First, experimental economists have had to fight hard to be accepted in a profession that for many years looked down on the relevance of experiments to the real world (a battle that is not over yet). Second, the primary goal of much experimental economics is theory *testing*, whereas, for psychologists, experiments are much more part of the process of *creating* theories.

Lest this latter statement be misinterpreted, let me rephrase as follows. Compared to economics, psychology does not have a strong theoretical paradigm and, consequently, the object of much experimentation is to provide evidence in support of theoretical notions favored by the author(s). For example, the structure of many papers is to reject conventional notions or models of behavior in favor of alternatives. As examples, consider the extensive literature in judgment on heuristics and biases and alternatives to expected utility as a rule for choice (Kahneman et al. 1982; Kahneman & Tversky 2000). In experimental economics, on the other hand, the object is typically to test implications of economic theory. For example, many experiments have tested predictions of auction theory, market clearing prices, institutional arrangements, and the like. Both disciplines, however, tend to publish papers that support the theories favored by those conducting the tests.

In my view, the main strength of the practice of experimental economics is the clarity about characteristics of people and tasks that are being investigated. At the end of the experiment, it is clear what level of generality of results the authors are claiming. Psychological experiments, on the other hand, are often distinguished by lack of clarity about the kinds of people and tasks to which the results might be generalized. Instead, the practice is to report results that typically have high internal validity, and then to let the reader surmise as to how such results might be generalized.

For example, following the early work on heuristics and biases, many people rushed to conclusions about cognitive skills based on studies that looked only at discrete units of behavior. And yet, when the same response tendencies were examined in terms of environments that allowed corrective feedback, a different picture emerged (Hogarth 1981). My point is not that the initial studies were flawed; it is that insufficient attention was paid to specifying the conditions (types of subjects and environments) to which the results applied. Indeed, there is ample evidence in the real world that people can exhibit significant biases or make errors in judgment in important discrete or one-shot situations. For instance, the result of the latest Presidential election in the United States may well have been determined by a "framing" effect induced by the so-called "butterfly ballots." A significant number of people misread the ballot and failed to vote for the candidate of their choice.

Parenthetically, economists often deride the outcomes of psychological experiments because they point out that the behaviors exhibited would not exist under market conditions. At one level, they may be correct. But the criticism misses the point. As just noted above, not all significant behavior takes place in markets and, even in markets, there is room for many different types of behavior. Thus, whereas experimental economists are effective in conducting experiments that can, in principle, be generalized, they are also considering only a small subset of human behavior.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

In short, I do not follow the view that psychologists should adopt all the practices of experimental economists. I am unconvinced, for example, of the need to remunerate subjects by performance (see Camerer & Hogarth 1999) or to repeat trials (see comments above on one-shot situations). Moreover, studying the behavior of inexperienced decision makers is important in its own right. On the other hand, I do think that experimental practice in psychology would be improved immensely if we thought more carefully about characteristics of people and tasks that allow general statements to be made. In this regard – and contrary to the emphasis of H&O – I believe that instead of considering the experiment as the unit of analysis, we should think more in terms of *programs* of research. Results found in the laboratory may have high internal validity; however, can studies also be done outside the laboratory that could shed light on external validity and limiting conditions? Conducting programs that involve both field and laboratory studies could help psychologists understand what their findings mean (for some good examples, see Kahneman & Tversky 2000).

Varying the scale of financial incentives under real and hypothetical conditions

Charles A. Holt[a] and Susan K. Laury[b]

[a]*Department of Economics, University of Virginia, Charlottesville, VA 22903;* [b]*Department of Economics, Georgia State University, Atlanta, GA 30303-3083.* **cah2k@virginia.edu slaury@gsu.edu www.people.virginia.edu/~cah2k www.gsu.edu/~ecoskl**

Abstract: The use of high hypothetical payoffs has been justified by the realism and relevance of large monetary consequences and by the impracticality of making high cash payments. We argue that subjects may not be able to imagine how they would behave in high payoff situations.

A psychologist, Sidney Siegel, has been largely responsible for establishing the procedural standards used in economics experiments. His work used salient financial incentives (e.g., Siegel & Fouraker 1960), appropriate non-parametric statistics (Siegel 1956), and clear scripts for subjects. In one classic study, the conclusions of the previous twenty years of probability matching experiments were reversed by using small financial incentives (Siegel & Goldstein 1959). In order to address important issues that involve large sums of money, some psychologists advocate using high hypothetical payoffs. For example, Kahneman and Tversky (1979, p. 265) recommend the method of large hypothetical payoffs over the "contrived gambles for small stakes" that are typical in economics experiments:

> The use of the method relies on the assumption that people often know how they would behave in actual situations of choice, and on the further assumption that the subjects have no special reason to disguise their true preferences. (Kahneman & Tversky 1979)

A comment that we often hear at interdisciplinary conferences is: "If it were just a matter of paying our subjects pennies, we would pay them, but we are interested in major decisions." Hertwig and Ortmann (H&O) do not evaluate this economic realism justification for high hypothetical payoffs, nor do the studies they discuss deal directly with the issue of whether the scale of payoffs matters, either in real or hypothetical payment situations.

To address payoff scale effects, we consider a lottery choice situation in which there is some evidence that the nature of payoffs does not matter. Tversky and Kahneman (1992, p. 315), for example, note that in choices between risky prospects, "we did not find much difference between subjects who were paid a flat fee and subjects whose payoffs were contingent on their decisions." We also observed this type of payoff invariance for choices between lotteries of the form:

> Option A (safe): probability p of \$2.00 and 1-p of \$1.60,
> Option B (risky): probability p of \$3.85 and 1-p of \$0.10,

where p is varied from 0.1 in decision 1, to 0.2 in decision 2, . . . to 1.0 in decision 10. As reported in Holt and Laury (2000), each person made all ten decisions, with one selected at random *ex post facto* to determine earnings on the basis of the subject's choice for that decision and the random outcome. In total, 93 subjects made these 10 choices, followed by a menu of hypothetical choices with all payoffs scaled up by a factor of 20 (payoffs of \$40 or \$32 for the safe option versus \$77 or \$2 for the risky option). After the determination of these "high" hypothetical earnings, the same subjects were asked to make the same 10 choices with the high payoffs being paid in cash.

The results are summarized in Figure 1, which graphs the percentage of safe choices in each decision. It is straightforward to

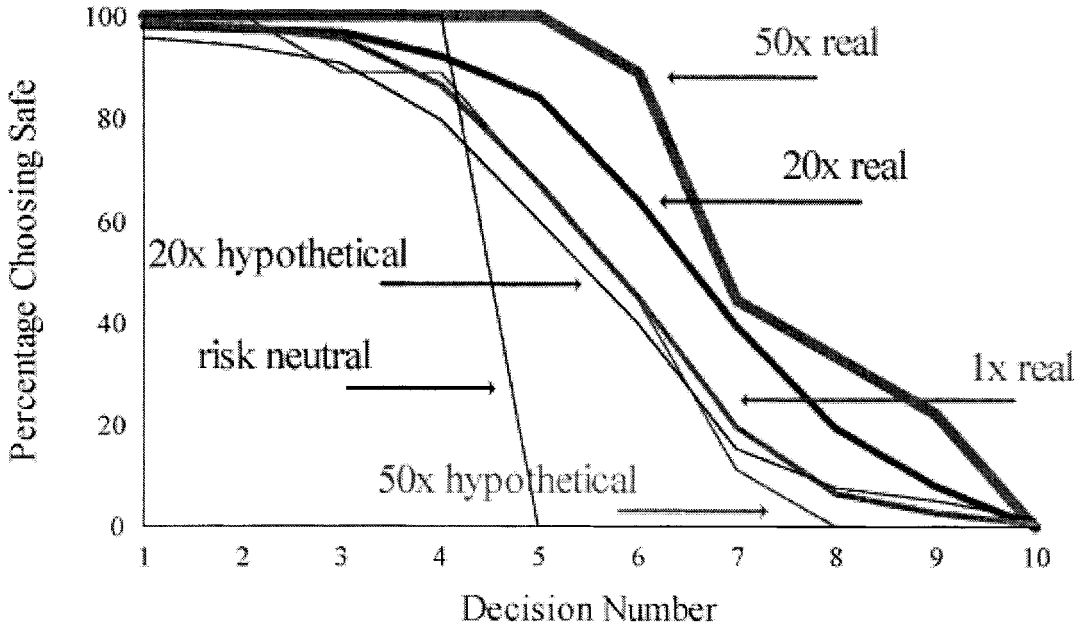


Figure 1 (Holt & Laury). Percentage of safe choices under real and hypothetical conditions (1x, 20x, and 50x payoff scale).

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

verify that a risk neutral person would choose the safe option in the first four decisions and switch to the risky option as soon as the probability of the high payoff exceeds 0.4, as shown by the straight thin line labeled "risk neutral." The choice percentages for the low real payoff condition, shown by the line labeled "1x real," are generally to the right of the risk neutral line, indicating some risk aversion. The choice percentages for the high hypothetical condition are represented by the thin line labeled "20x hypothetical," which is quite close to the line for the low real payoff condition, and indeed the difference is not statistically significant. In contrast, the thick line labeled "20x real" for high real payoffs lies to the right, indicating a sharp increase in observed risk aversion when these payoffs are actually made. In addition, we scaled payoffs up from the original level by a factor of 50 ($100 or $80 for the safe option, versus $192.50 or $5 for the risky option). The data for the 9 subjects in this rather expensive treatment are shown by the thick "50x real" line, which indicates a further increase in risk aversion. In fact, none of the subjects were risk neutral or risk seeking for this very high treatment. Notice however, no such scale effect is seen in the hypothetical choices; the thin "50x hypothetical" line is quite close to the "20x hypothetical" and "low real" lines.

These results illustrate how the use of low monetary incentives may not matter, but it does not follow that using high hypothetical incentives is a way to investigate behavior in high-stakes economic choices. In addition, the use of low or hypothetical payoffs may be misleading in that biases and non-economic factors may have an impact that is not predictive of their importance in significant economic choices. For example, women were significantly more risk averse than men in the low-payoff condition, but this gender effect vanished in the high-real-payoff condition.

Some data patterns for experiments with hypothetical payments may not be robust, which may cause journal editors to be hesitant about accepting results of such studies. For example, when the above options "A" and "B" are doubled and reflected around zero, we do not observe strong reflection (risk aversion for gains and risk seeking for losses) with real money payments although such reflection occurred in over half of the cases when gains and losses were hypothetical (Laury & Holt 2000). In the real payoff condition, only about 10% of the subjects reflected, and another 5% reflected in the wrong direction (risk seeking for gains and risk aversion for losses). These choice patterns are sharply different from the widely cited results of Kahneman and Tversky (1979) using hypothetical payoffs over different lottery pairs.

In the absence of a widely accepted theory of when financial incentives matter, we believe that performance-based payments should be used in economics experiments, even in seemingly similar situations where no effect has been reported in past studies.

Variability is not uniformly bad: The practices of psychologists generate research questions

Scott A. Huettel[a] and Gregory Lockhead[b]

[a]*Brain Imaging and Analysis Center,* [a,b]*Department of Psychology, Experimental, Duke University, Durham, NC 27708.*
{huettel; lockhead}@duke.edu

Abstract: The practices of economists increase experimental reproducibility relative to those of selected psychologists but should not be universally adopted. Procedures criticized by Hertwig and Ortmann as producing variable data are valuable, instead, for generating questions. The procedure of choice should depend on the theoretical goal: measure a known factor or learn what factors are important and need to be measured.

According to Hertwig and Ortmann (H&O), experimental procedures adopted by economists allow for reproducibility of results, facilitate compliance with instructions, and convey experimental goals faithfully, and psychologists ought to follow these procedures but do not. These are worthy aims but we disagree that experimenters always should specify precisely the experimental setting for the purpose of reducing participant uncertainty. These proposed methods ought to be rigorously upheld when the research question is well understood, as is done in most areas of psychology. But when the question is not understood, then precisely conveyed experimental goals might provide reliable data for the experimenter's task, but might not usefully facilitate learning what people ordinarily do in free situations. It is not that the methods of psychologists and those of economists conflict, but that the economic studies and the particular psychology studies that were selected for consideration by H&O have different goals. When one wants to answer a well-defined question, then strict procedures as described by H&O ought to be used. Indeed, these procedures do not belong to economics but are followed by all scientists, including most psychologists. But when the situation is not well defined and the experimenter wants to learn what might lead to productive research questions, then procedures that allow subjects to interpret the situation can be valuable.

For their analysis, the authors focus upon four practices of (some) psychologists: lack of script provision, failure to repeat trials, absence of subject payment, and inclusion of deception. They summarize use of these practices fairly, and we agree that such practices reduce replicability of results. These practices reflect a methodological subset of all practices in psychology. Other, and perhaps more important, theoretical practices include generation of new hypotheses, conduct of exploratory analyses on novel experimental designs, and psychophysical assessments of mental states. For just one example, consider the issue of understanding how a choice by one person is affected by his or her expectations about the choices that might be made by other people. To address this, we employed a cooperative game (known as "Stag Hunt" or "Wolf's Dilemma") where every member of a group of players gains a large reward only if everyone cooperates, and where individuals who defect gain a small reward no matter what other people do (Huettel & Lockhead 2000). Trials were repeated many times, payment was in lottery tickets based on performance, and no deception was introduced. However, we violated the first axis of H&O by placing subjects in competition with computer opponents, and having them report their assessments of their opponents' likely choices before each trial. Relaxation of this proscription allowed us to specify what the opponents (i.e., the previously programmed computer responses) would do across trials, which provided variability in subjects' expectations of their opponents' behavior.

One result from that study shows the relation between the proportion of subjects who cooperated and the proportion of opponents the subjects expected to cooperate (Fig. 1). Each data point represents a different sequence of preceding events. Across these sequences, when subjects expected no one to cooperate on the next trial, they too did not cooperate, as visible at the lower left of Figure 1; when subjects expected opponents to cooperate, they also cooperated, as shown at the upper right of Figure 1; and the proportion of subjects cooperating grows between these extremes. The solid line in Figure 1 is not a best-fit curve. Rather, it is the result of the following calculation: convert the value on the x-axis to the probability that one or more opponents will defect, multiply this probability by the cooperate/defect payoff ratio, and divide the result by the sum of the payoffs. The apparent close fit of this function to the data indicates that choices in this experiment are largely explained by subjects' expectations of what opponents would do: When the sequence of prior events leads the subject to expect opponents to cooperate, then the subject is likely to cooperate. Thus, this psychological measure of expectation predicts this behavioral/economic measure of choice. Without the additional variability in responses that was introduced by having different computer strategies, this relation could not have been readily identified.

The data point that is most deviant from this explanatory function (at $x = 0.78$, $y = 0.35$) provides further evidence for the im-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

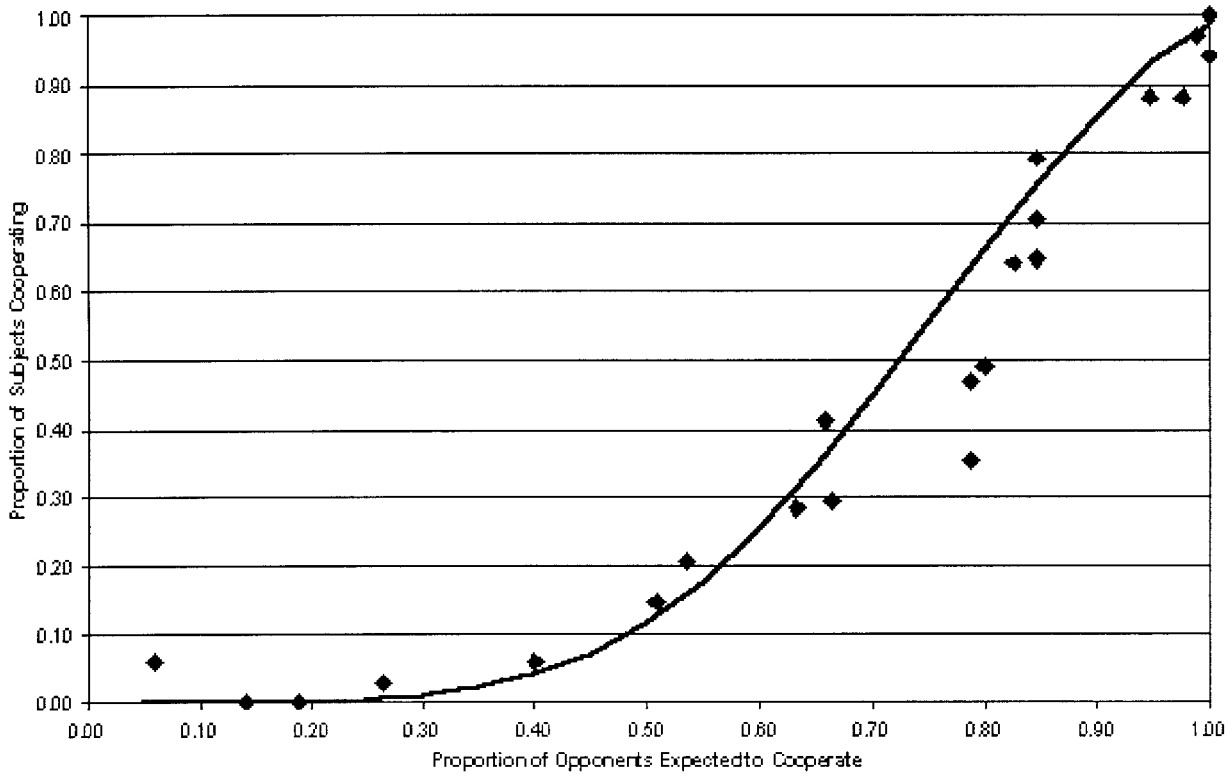


Figure 1 (Huettel & Lockhead). The relation between subjects' choices (y-axis) and their expectations of others' choices (x-axis) in a cooperative game. The line indicates the expected utility function translating the psychological measure of expectation to the behavioral measure of choice behavior, given the parameters used in this game.

portance of looking for new questions. This point represents sequences of three trials in which there were three defections and the same opponent defected on each trial. This caused a lower than predicted rate of cooperation. When, instead, a different opponent defected on each of the three prior trials, and so there again were three prior defections, then the proportion of subject cooperation was much higher (Huettel & Lockhead 2000, Figs. 4, 5). Thus, choices are not only sensitive to the overall likelihood of cooperation but, further, people attend to failures of cooperation by individuals.

We do not suggest that results such as these fully describe how choices depend on expectations of the behavior of others. However, we do suggest that topics uncovered by such studies, topics such as the sequence of events, the regularity of individual opponent's behaviors across trials, and the number of opponents involved, then can be brought under the tighter controls proposed by H&O and used routinely in most psychology and other laboratories. Perhaps, the particular psychologists addressed by H&O differ from these economists in what they are attempting to understand about decision behavior. Hertwig and Ortmann wish to establish how specific, known factors contribute to decision behavior, while these psychologists may wish to learn what additional factors might contribute to choice. This latter is a more imprecise approach and following it exclusively can lead to the sort of research circularity that plagues much of the behavioral sciences. Nevertheless, failure to introduce variability into experimental design risks missing unexpected aspects of decision making, and often leads to continuing measurement on factors that are not fundamental to the behaviors being studied but are contaminated with those factors (Lockhead 1992). The balance between restricted experimental designs, which allow reproducibility and hypothesis testing, and exploratory experimental designs, which may provide new insights into phenomena, is not particular to decision-making research. In general, no disciplines should be delineated by the research practices they employ and any scientist, whether psychologist or economist or something else, should wield tools as needed. Indeed, the two sets of techniques characterized by H&O as reflecting psychological and economic practices have different strengths, depending on the theoretical approach of the research. We suggest that the difference between the particular psychologists and economists considered by H&O lies not in their research practices but in their research questions.

Why use real and hypothetical payoffs?

Anton Kühberger

Department of Psychology, University of Salzburg, A-5020 Salzburg, Austria.
anton.kuehberger@sbg.ac.at
www.sbg.ac.at/psy/people/kuehberger.html

Abstract: Decision making can be studied using hypothetical payoffs because it is hypothetical to its very core. However, the core process can be influenced by contextual features. As there is no theory for these contextual features, a "do-it-both-ways" rule amounts to a waste of money. If we had such a theory, doing it both ways would be unnecessary.

In their discussion of the role of financial incentives, Hertwig and Ortmann (H&O) correctly point out that the use of financial incentives is obligatory in economics, but the exception in behavioral decision making. Further, they tend to side with the economists by concluding that, in general, financial incentives affect performance for the better. They report the results of a meta-analysis on the use of financial incentives in judgment and decision-making and propose that psychologists in behavioral decision making consider using financial incentives. I do not agree with their evaluation of the usefulness of financial incentives. First,

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

there is a deeper reason than simply the cost for the pervasive use of hypothetical payoffs in behavioral decision theory. Why do psychologists believe that the study of reactions in imagined situations is a legitimate means of studying real decision behavior? To be sure, in other areas of psychology (for instance, in psychophysics), such methods would be considered extremely questionable if not absurd. The reason is that decision making – rather than, for example, perception – is hypothetical at its very core. When making a decision, we anticipate hypothetical states of the world, we consider events that could or could not obtain, we consider feelings that we do not have yet. At the time of decision, none of these outcomes, events, or feelings, is real, but all are hypothetical. That is, in essence, decision making consists of the manipulation of hypothetical mental contents. Thus, decision researchers have some justification in assuming that people's real decisions can profitably be investigated by asking them to make hypothetical decisions. However, this does not necessarily mean that all real decisions can be studied hypothetically. The problem is that the core process of decision making, although hypothetical, may be influenced by the context, for instance, by the importance of the decision, or in other words, by the incentives involved. But, the crux is that we have little knowledge concerning the context conditions that change the core process reliably. For instance, we have shown (Kühberger et al. 2000) that the framing effect exists for hypothetical, as well as for real moderate to high incentives (approx. $50), but we still don't know whether it exists for real decisions involving hundreds of human lives. Without a theory of why (some) hypothetical decisions match real decisions, the whole advantage of hypothetical decision research is defeated because for each question to be answered by a hypothetical decision experiment the results have to be validated by an accompanying real decision experiment. However, if the real thing is done, there is no further need for a hypothetical one.

H&O do touch upon this problem in their discussion of the reasons for mixed results, where they report two possible explanations for these mixed results. However, the "lack of payoff dominance" explanation fails because of its too narrow-minded focus on incentives alone (see Camerer & Hogarth 1999; Kühberger et al. 2000). The "multiple and contradictory norms" reason, on the other hand, shows the vulnerability of experimentation based solely on the incentive structure to be of limited external validity (Loewenstein 1999). What is needed is a theory of when and why real and hypothetical decisions will match, and Camerer and Hogarth (1999) present an important step for such a theory. As long as we lack such a theory H&O's proposal of a "do-it-both-ways" rule amounts to a waste of money; as soon as we have a theory, there is no need for doing it both ways.

Are experimental economists behaviorists and is behaviorism for the birds?

Robert Kurzban

Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125. **rkurzban@hotmail.com**

Abstract: Methods in experimental economics are reminiscent of the methods employed by behaviorists in the first half of the twentieth century. Empirical and conceptual progress led the field of psychology away from the principles of behaviorism, and experimental economists should consider whether the criticisms leveled against behaviorists might apply equally to them.

Hertwig and Ortmann's (H&O's) discussion of the dimensions of methodological practices in experimental economics is eerily evocative of the "dark days of behaviorism" that characterized the field of psychology in the first part of the twentieth century. Like some experimental economists, Pavlov, Skinner, Thorndike, and their intellectual descendants, who used "performance-based incentives" (often in the form of Purina Laboratory Chow), never deceived their subjects, and generally used a sizable number of trials, often numbering in the hundreds.

More generally, the stripped-down experimental context of the typical behaviorist experiment, the soundproofed bare box with lever and pellet-dispenser, sounds a great deal like the idealized context endorsed by experimental economists. Behaviorists believed for their purposes that "the best thing to do is bring the pigeon or rat into an artificial, laboratory environment that prevents species- and situation-specific behavior patterns from exerting their influence." (Schwartz 1989, p. 44). Similarly, H&O suggest that, "in an attempt to control home-grown priors (i.e., beliefs and attitudes that participants bring into the experiment), the scripts provided by economists are typically as content-free as possible" (n. 4).

These methodological similarities reflect deeper similarities in theoretical orientation. Economists and behaviorists model the organisms they study as having a set of relatively simple preferences (e.g., more money or food is better than less; less effort or electrical shock is better than more) and a set of rules to learn how to satisfy these preferences in the environment in which the organism finds itself. The precise specification of these rules is taken to be a subject, or perhaps *the* subject, of empirical investigation. Given this view of the organism, it makes sense to use incentives, repetition, and highly abstract environments; features that allow an uncluttered look at the learning mechanisms.

However, the adoption by some economists of the methodological practices of behaviorism might strike some psychologists as somewhat odd. Theoretical and empirical progress that began in the sixties led psychology as a field to largely abandon behaviorism and its methods. Blows to behaviorism came from several sources. Garcia and Koelling (1966) showed that organisms must come to the experimental situation with privileged hypotheses, rather than blank slates open to any association between stimulus and reinforcement. Breland and Breland (1961) showed that there are important limits on how reinforcement can shape the behavior of organisms. Finally, Premack and Woodruff (1978), building on Köhler's well-known experiments demonstrating insight learning in chimpanzees, showed a chimp could solve complex problems without any prior reinforcement or conditioning of the appropriate behavior.

In short, these experimental findings and others like them brought the curtain down on behaviorism as the theoretical framework in psychology because it had insufficient theoretical power in describing the way organisms actually behaved. Behaviorism not only failed to account for numerous elements of organisms' behavioral repertoires, but also proved inadequate even in its area of strength, describing how organisms learned. Behaviorist principles gave no account of the large number of cases in which learning took place without reinforcement, including phenomena such as "latent" learning, imitation, and insight. More important still were the conceptual insights of the cognitive revolution, including Chomsky's (1975) work showing that conditioning alone could not, in principle, explain how children learned language. Later work in related fields reinforced this conclusion, showing that behaviorism's associationist learning principles also could not account for phenomena in non-linguistic domains.

Associative learning: The last resort? An alternative to the behaviorist view is that cognition consists of a set of domain-specific computational devices designed to solve the adaptive problems associated with the organism's natural history. These systems are deployed in a situation-specific way, and are powerful because they can make certain assumptions about the environment the organism will face (Tooby & Cosmides 1992). However, because evolution cannot anticipate every possible contingency an organism might face, one possibility is that natural selection has endowed many organisms with a set of relatively weak, domain-general learning systems that operate on the principles of pleasure and pain, generating the law of effect and the other elements of behaviorism. These systems generate the well-known learning

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

curves observed in the multitude of studies that used highly stylized and content-free stimuli.

Thus, perhaps these learning mechanisms should be construed as back-up systems that the organism engages when it is confused by evolutionarily novel stimuli (such as electric shocks), used as a last resort when its domain-specific systems cannot be applied. It can be argued, therefore, that behaviorist labs were illuminating this backup learning system. In contrast, experiments such as those by Garcia and Koelling were illuminating what would be construed by some as the more interesting part of rat psychology, its "home-grown priors" that allow it to navigate the world successfully. While the behaviorists expanded our understanding of the learning system that had evolved in rats to solve problems they had never encountered during their evolutionary history, Garcia and those that came after him dissected the evolved content-rich cognitive architecture of the rat.

If this analysis is correct, behaviorist-like experiments in humans might also be engaging content-independent mechanisms that exist because evolution cannot foresee all possible contingencies. Experiments with stripped-down contexts, monetary rewards, and repeated trials might indeed allow us to view these learning systems in sharp relief, as illustrated in recent work detailing models of "fictitious play," "quantal response," and so forth (e.g., Camerer & Ho 1999). However, it is worth considering what the results of these experiments are really telling us about human psychology, and what they might be telling us about the last resort of a confused organism.

Other scientific purposes, other methodological ways

Marie-Paule Lecoutre and Bruno Lecoutre

ERIS, Laboratoire de Psychologie et Mathématiques Raphaël Salem, C.N.R.S. et Université de Rouen, 76821, Mont-Saint-Aignan Cedex, France
{marie-paule.lecoutre; bruno.lecoutre}@univ-rouen.fr
www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm

Abstract: Hertwig and Ortmann have made a laudable effort to bring together experimental practices in economics and in psychology. Unfortunately, they ignore one of the primary objectives of psychological research, which is an analytic description of general cognitive processes. Among experimental practices in probability judgment tasks they discussed, we will focus hereafter on *enactment of scripts* and *repetition of trials.*

While economists run experiments in a normative perspective, namely, to test decision-theoretic or game-theoretic models, most cognitive psychologists have no such priority motivation. Indeed a primary objective of psychological research in probability judgment situations is an analytic description of general cognitive processes involved in a whole class of tasks. Of course normative models have a role to play in defining and constructing situations of interest. Furthermore, linking experimental findings to the "optimal" behavior in a given task should also contribute to elaborate formal descriptive models of cognitive judgments. However, to list "errors" and deviations from a priori models is clearly insufficient.

Cognitive psychologists need another approach in order to investigate spontaneous cognitive processes and to provide evidence of a number of fundamental probabilistic intuitions. A further aim is to reveal some internal coherence in these processes. These scientific purposes call for specific methodological ways and experimental practices. In particular, a constant concern of cognitive psychologists should be to avoid as much as possible experimental situations inducing stereotypical or learned answers, reflecting subjects' theoretical knowledge (for example, in probability theory) or experience more than their own opinions and judgments. An application of such an approach to statistical inference situations can be found in Lecoutre (2000) and Lecoutre et al. (2001).

Moreover, rather than to repeat trials of one particular task with a precisely defined "script," it is desirable to vary the situations for characterizing the best conditions under which the appropriate cognitive processes are activated. Only such a variability can allow us to characterize processes generally enough to be transferable to a whole class of situations. In this approach, with situations which the subjects are led to construct themselves, the adequate representations are increasingly privileged. Such an active construction appears to be a determining factor in the stabilization of these representations. A recent statement by Fischbein and Schnarch (1997) refers to this approach: "If students can learn to analyze the causes of the conflicts and mistakes, they may be able to overcome them and attain a genuine probabilistic way of thinking." Furthermore, many recent research programs in probabilistic and statistical education emphasize that it is important for students to construct their own knowledge and develop probabilistic and statistical concepts through the use of active learning.

Finally, one can get worried about the generalisability of the results obtained from a precisely defined script when it is well known from the analogical transfer literature how much "cover stories" or semantic contexts can affect transfer. The issue of transfer of learning from one situation to another is of perennial interest to psychologists. Results could be interpreted within the framework of a general mechanism which is increasingly recognized as playing an important part in cognitive activity: analogical processing. A lot of experimental evidence in psychology has shown that the frequency of the use of analogy is due to its heuristic and economical nature which allows people to make "mental leaps" (Holyoak & Thagard 1995) between different domains, and to interpret a new situation in terms that transform the newness into a well-known situation. Usually analogical processing is studied in an experimental paradigm in which a "source" situation (solutions in problem-solving or a set of knowledge in a domain) is taught to the participants before testing their behavior within a "target" situation (the new problem or new domain). It is commonly accepted that one may describe this process as a comparison mechanism which allows people to recognize and infer similarities between situations. When a subject has to solve a new situation in which no source situation is given, he uses his own source analogue evoked or activated by the (semantic) context of the new situation.

Much recent research focusing on the conditions under which transfer occurs or fails to occur between two domains shows that transfer often fails to occur. Indeed, subjects most often have difficulty in using a source problem to solve either a close or distinct variant of this problem. In this context Robertson (2000) indicates the beneficial effect of providing an explanation at a level of generalisability sufficient to allow the subjects to adapt the procedure to suit the target problem. Even if the perspective of the experiment is "to give participants a chance to adapt to the environment, that is, to accrue experience with the experimental setting and procedure," such an approach can be an attractive alternative to the use of repeated trials of the same particular situation. Rather than acclimatizing subjects to a specific task with a precisely defined script, we may attempt to act upon the cognitive representations and to give subjects the opportunity to learn processes sufficiently general to be transferable to a whole class of situations.

In partial defense of softness

Daniel S. Levine

Department of Psychology, University of Texas at Arlington, Arlington, TX 76019-0528. **levine@uta.edu**
www.uta.edu/psychology/faculty/levine

Abstract: The authors wish that the psychology of human decision making should borrow methodological rigor from economics. However, unless economics also borrows from psychology this poses a danger of overly limiting the phenomena studied. In fact, an expanded economic theory

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

should be sought that is based in psychology (and ultimately neuroscience) and encompasses both rational and irrational aspects of decision making.

Hertwig and Ortmann's (H&O's) goal is to increase the amount of methodological rigor, and therefore robustness of results, in psychology, particularly the psychology of human decision making. In some aspects of this science they see a greater rigor coming from experimental economics and seek to expand the influence of economics on the design of experiments in psychology. Certainly a greater and more cooperative dialogue between the two fields will provide many benefits to both. Yet if this influence is only in one direction, there is a danger that the phenomena to be studied will be overly limited. There also needs to be a continuing and expanding influence of psychology on the development of economic theory.

In particular, we need to work toward a theoretical understanding of human behavior which encompasses both optimal and nonoptimal behavior; both one-shot and repeatable behavioral phenomena; and both heuristic-driven and rationally planned decision making. All these types of behavior are part of the human experience, and all are rooted in the way our brains are organized and learn from experience. Hence, psychology will increasingly build its theoretical foundations under the influence of neuroscience (see Gazzaniga 1995), neural network theory (see Levine 2000; Martindale 1991; Reed & Miller 1998), and dynamical systems (see Abraham et al. 1990). Since economic behavior is subject to the same influences as other types of behavior, these changing foundations of theoretical psychology are likely to have increasing influence in economics, not only on the design of economic experiments but on the foundations of economic theory itself.

For example, the authors cite much evidence that repetition of trials, performance feedback, and financial incentives frequently reduce many of the forms of decision making irrationality that Tversky and Kahneman have popularized, such as preference reversals, base rate neglect, violations of Bayes's rule, and the use of heuristics such as representativeness. Yet these results argue not for the unimportance of such irrationalities but for their context dependence. This is because a large number of economic decisions are made under time pressure, without the benefit of feedback, or under confusing and uncertain incentives, as pointed out by Thaler (1987) whom the authors cite. Such hurried or one-shot decisions are made both by professional investors (see Shefrin 2000) and by average consumers; an example of the latter is the preference reversal between Old Coke and New Coke going from controlled taste tests to the actual market (which has been modeled in a neural network by Leven & Levine 1996). In particular, the Coke example addresses the observation of Roth (1995), cited by the authors, that "the question of actual versus hypothetical choices has become one of the fault lines that have come to distinguish experiments published in the economics journals from those published in psychology journals" (p. 86).

I agree with H&O that psychological data have frequently been used to justify a form of cynicism about human rationality in economic decisions, and that this cynicism is misguided. As they say in the target article, "People's performance in early trials . . . does not necessarily reflect their reasoning competence in later trials." Indeed, the "irrationality" results of Tversky and Kahneman (1974; 1981) and many others can be used to argue that people have stronger reasoning capacities than they appear to show under time pressure or other forms of stress (see Leven & Levine 1995, for such an argument). The often cited work of Cosmides and Tooby (1996) showing that humans are good at Bayesian reasoning *if the context is structured to encourage it* points in the same direction.

Yet this argument can only be made effectively within an economic theory that is psychologically based; that is, one that incorporates both rational and nonrational influences, including the influences of emotion, habit, and novelty. This needs to be a theory that transcends the cultural assumptions that reason is superior to emotion and encompasses both the destructive and the productive influences of emotion and intuition. It will be a far different kind of "economic theory" than that the authors discuss in the target article: "most economics experiments test economic theory, which provides a comparatively unified framework built on maximization assumptions (of utility, profit, revenue, etc.) and defines standards of optimal behavior." It will, as per H&O's prescriptions, not discard repeatable economic phenomena or those subject to feedback but embed them in a larger framework built on knowledge of the human brain. It will subsume current mainstream economic theory much as relativistic mechanics subsumes Newtonian mechanics.

What forms of experiments make the best tests of such a psychologically based economic theory? The current mainstream type of economics experiments, involving financial incentives and scripts, certainly form part of the data base required. Yet so do other types of experiments that involve studying the way people actually make economic decisions, without prompting and without feedback. Clearly, as the authors suggest, the two disciplines have much to learn from each other. Yet the polarization between "soft," "psychological" and "rigorous," "economic" approaches is just as silly as other polarizations in the study of behavior, such as nature versus nurture. Only a dynamical scientific theory that bridges both will yield answers that are genuinely predictive and applicable to real-life situations.

We should not impose narrow restrictions on psychological methods

Michael Maratsos

Institute of Child Development, University of Minnesota, Minneapolis, MN, 55414. **marat001@tc.umn.edu** **www.umn.edu**

Abstract: Hertwig and Ortmann suggest greater standardization of procedures in experimental psychology to help with problems of replicability and consistency of findings. It is argued that, (a) this view is inconsistent with their other interesting proposals, and (b) heterogeneity of method is appropriate in psychology.

Hertwig and Ortmann (H&O) skilfully argue for a number of specific proposals: in decision-making studies, they suggest more use of high-practice, high-knowledge studies; more role-taking studies; a greater specification of theoretical targets in the real-life situations used for these variations. More generally, they propose a minimal use of deception.

But throughout the course of their argument, they also suggest that experimental psychologists need to do as experimental economists have done: that is, narrow the variations in their procedures. This argument appears to reach far beyond studies of decision-making, as general remarks about "soft psychology," clinical psychology, and highly general figures in the history of psychology are cited.

Their own suggestions about decision-making studies, however, seem to contradict this goal. Mostly what they suggest would lead to *greater* variety in psychological work, compared to current practices: more role-playing studies, more use of contingent payment, more use of high-practice, repeated-trial, high-knowledge conditions, more tailoring of procedures to carefully thought-out real world target behavior and theory domains. It is true they suggest that contingent versus noncontingent payment be adopted as a standard independent variable. Their suggestion that studies be carefully tailored to their real-world target behaviors implies that experimenters only use one procedure if it is relevant. They do not strictly propose adding procedures not already in (rare) use. But making rarely-used procedures relatively more common does increase the variety of *standard* and therefore *salient* approaches. Hence, I interpret their proposals as increasing variety, not decreasing it.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

The deeper issue is whether standard and narrow procedure really is the right way for psychological investigation. It probably is not. Ernst Mayr (1988) in *The history of biological thought* argues that in biological systems, attempts to treat biology and evolution as though they were physics generally fail. Basic physics emphasizes a very few principles operating in standard ways. But the major characteristics of biological systems and development, and evolution itself, arise from the diversity of adaptive mechanisms and circumstances. A few general principles, like the nature of DNA, have general importance. But breaking down how things work in practice in species' physiology, neurology, behavior, and evolutionary development, typically require awareness of, and attention to, system and behavior diversity. Attempts to operate like physics typically divert thought from this diversity, and therefore take a wrong track.

Human psychology, a biological system, shows this same characteristic of diversity of behaviors and systems in diverse situations. This requires from investigators a willingness to be flexible in pursuing knowledge of different types of psychological behaviors and systems. Attempting to put everything into the same narrow methodology and procedure is essentially incorrect, though it may be correct in highly constrained subjects of study (like psychophysics).

Then, it is complained that findings are difficult to replicate, or vanish with minor variation. Here, first of all one may hope that the study of why these variations occur can itself contribute more valuable knowledge. Further, what if this changeability represents human functioning as it actually is? There is, in fact, a converse danger that in narrowing how we study things to a very constrained set of procedures, we may construct a "fairy tale" depiction of human life, one more "consistent" and replicable than human life really is across its variety of situations. American behaviorists in effect constructed such a fairy tale for themselves, which eventually had to be broken down. If diverse adaptive mechanisms, responsive to variations in situations, is a key aspect of human life (as seems likely), simply dispensing with anything that relieves us of this central – if investigatively annoying – fact, is contrary to scientific truth. And indeed there are good reasons that "soft" areas of psychology, typically relevant to behaviors and beliefs that are the most responsive to variations in culture and situation, would show this characteristic most of all.[1]

Obviously the above arguments may be negotiated on a case-by-case basis, to make our work as scientists more practicable. (More space would permit examples here.) But again, if making life more "practicable" becomes our central goal, the induced distortions will probably make us in some deeper sense more ignorant, not less, given the human psychology we are dealing with. As noted above, H&O really seek to introduce more diversity of standard practice, not less, even in the limited domain of decision behavior. In this, their basic instincts about increasing knowledge fortunately predominate over their desire for a neater methodological world.

NOTE

1. And indeed, even worse would be the constraint of psychological method itself to "true" experimental method, as many psychologists seem to desire. Given all the central matters that cannot at all be studied in laboratories for ethical and resource reasons, this is absurd. In the scientific laboratory, can we give people conditions traumatic enough to elicit denial or repression? Can we make up new cultures, lived in for years, to see how people's behavior develops? Rhetorical questions.

Choice output and choice processing: An analogy to similarity

Arthur B. Markman

Department of Psychology, University of Texas, Austin, TX 78712.
markman@psy.utexas.edu
www.psy.utexas.edu/psy/faculty/markman/index.html

Abstract: The target article suggests that many practices of experimental economists are preferable to those used by psychologists studying judgment and decision making. The advantages of the psychological approach become clear when the focus of research shifts from choice output to choice processes. I illustrate this point with an example from research on similarity comparisons.

Output and processing. The target article suggests that the experimental procedures followed by experimental economists are often to be preferred to those used by psychologists. These procedures are likely to decrease variability in choices, to lead to highly motivated participants, and to minimize participants' suspicions about the experimental situation. The authors suggest that many conflicting findings in the judgment and decision making literature may be a function of differences in the experimental methods used by researchers.

One reason for the variability in studies of judgment and decision making is that studies focus on choice output rather than on choice processing. That is, the dependent measures of these studies consist primarily of choices of one of a set of options or judgments about some situation. Choice output is variable, because different participants are likely to adopt very different goals in an experimental situation. Unless the experimental situation is narrowly defined in a way that constrains people's goals (as in many studies in experimental economics), there is likely to be significant variation in performance across participants.

An alternative approach to decision making research focuses on choice processing (e.g., Payne et al. 1993). This approach assumes that variability in choices masks conformity at the level of the processes people use to make choices. To illustrate this point, I first draw a brief analogy with research on similarity comparisons. Then, I explore why techniques from experimental psychology are better suited to the study of choice processing than are methods from experimental economics.

The analogy with similarity. Until the early 1990s, the dominant method for studying similarity was to present people with pairs of items and ask them to assess their similarity through tasks like ratings. Models were then developed to account for these similarity ratings. These models typically involved multidimensional similarity spaces (e.g., Shepard 1962) or featural representations of objects (e.g., Tversky 1977).

A drawback of these models was that it was difficult to capture the effects of context on similarity comparisons. The prevalence of context effects in similarity judgments led Tversky (1977) to reject spatial models as good explanations for patterns of similarity judgments. A similar logic was used in later studies to create patterns of data that featural models could not explain (e.g., Goldstone et al. 1991).These context effects typically involve cases in which particular types of properties are more important to similarity judgments in some cases than in others.

Later research on similarity comparisons skirted the influence of context effects on similarity judgments by focusing on the processes people use to make comparisons rather than on the similarity judgments themselves (Gentner & Markman 1997; Medin et al. 1993). This work suggests that similarity comparisons involve a comparison of structured representations in which people seek matching relational structures in two domains.

The stability of the underlying processes has had two salutary effects on research in similarity. First, it has led to considerable agreement across research groups about the appropriate way to model similarity behavior (e.g., Falkenhainer et al. 1989; Hummel & Holyoak 1997). Second, it has led to the application of these ap-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

proaches to similarity to other cognitive processes that involve comparisons (e.g., Zhang & Markman 1998).

There is a similar shift occurring in the study of judgment and decision making. The discipline of consumer behavior focuses on the way people make choices of consumer products (Jacoby et al. 1998). Unlike experimental economics and behavioral decision theory, this research is not focused on issues of the optimality of choices (an output concern). Instead, consumer behavior explores the processes used to make choices and the way those processes are affected by task factors such as the information available and the level of involvement of the consumer.

Processing and methods. I contend that the methods of experimental psychology are more valuable for the study of choice processes than are the methods of experimental economics. To make this point, I take research on consumer behavior as a case study. First, I briefly consider the four aspects of experimental situations discussed in the target article. Then, I discuss some additional points not considered in the target article.

First, consumer behavior relies on participants' prior exposure to consumer products and choice to provide stability in the task environment. In a typical study, participants are given information about consumer products. For most participants in experimental research (e.g., college undergraduates), this is a familiar situation. In this familiar situation, there need not be extensive repetition of trials. Indeed, analyses of people's choices from UPC scanner data suggest that most repeated choices (in supermarkets) are simply repetitions of choices made previously in the same situation (Guidagni & Little 1983). Thus, there is little value in looking at repeated trials.

The issue of participant motivation is a variable that is explicitly considered by researchers in consumer behavior. While monetary payments are rarely used, manipulations of participant involvement may be instantiated through instructions or through individual differences in personality variables (e.g., Johar, 1995). Finally, because participants are placed in a familiar task environment in which they are evaluating consumer products, there is often less suspicion of the experimental situation than there would be in other experimental situations. Consistent with the suggestions in the target article, it is possible to carry out many studies in consumer behavior without deception.

Where experimental psychology has an advantage over experimental economics is in the wealth of techniques for studying online processing. Studies may explore a range of behaviors in addition to choices. For example, recall information about choice options and retrospective verbal protocols may be collected. More elaborate process tracing methods include the Mouselab paradigm (a computerized information board; see Payne et al. 1993) and eye tracking (Russo & Dosher 1983).

Finally, by exploring decision making as a process, people's choice behavior can be unified with the rest of psychology. Rather than treating decision making as a separate area of research, consumer behavior looks at the role of memory, comparison, categorization, and other fundamental cognitive processes on decision making. This approach holds promise to provide a more stable view of judgment and decision making behavior.

Participant skepticism: If you can't beat it, model it

Craig R. M. McKenzie and John T. Wixted

Department of Psychology, University of California, San Diego, La Jolla CA 92093-0109. **{cmckenzie; jwixted}@ucsd.edu**
www.psy.ucsd.edu/{~mckenzie; ~jwixted}

Abstract: For a variety of reasons, including the common use of deception in psychology experiments, participants often disbelieve experimenters' assertions about important task parameters. This can lead researchers to conclude incorrectly that participants are behaving nonnormatively. The problem can be overcome by deriving and testing normative models that do not assume full belief in key task parameters. A real experimental example is discussed.

Hertwig and Ortmann (H&O) raise several important issues that need to be considered by psychologists, especially those who study judgment and decision making. Our commentary focuses on the widespread use of deception in psychology experiments. H&O point out that suspicious participants behave differently from trusting ones. Not mentioned by the authors, however, is that this can be especially problematic in tasks that compare participants' behavior to a normative standard, like those often used in judgment and decision making, because differences between behavior and the normative response are often interpreted as errors. A critical assumption in attributing errors to participants is that they fully believe the assumptions underlying the purported normative response (e.g., that the options are truly mutually exclusive and exhaustive, or that the observations are truly randomly sampled). Skepticism about task parameters – skepticism that is justified, given that participants are often deceived – can lead to a different normative response than assumed by the experimenters. This problem can be overcome, however, by deriving and testing normative models that do not assume full belief in important task parameters. We provide a real experimental example to illustrate our point.

Recently, we have examined the relationship between confidence in two widely used tasks in psychology: Yes/no and forced-choice tasks (McKenzie et al. 2000; 2001). Both tasks are routinely used in studies of perception, categorization, memory, and judgment and decision making. For example, imagine reporting confidence in the truth of two general knowledge statements presented sequentially: (A) The population of the U.S. is greater than 265 million. (B) Sophocles was born before Socrates. Assume that you are 80% confident that A is true and 40% confident that B is true. You are subsequently informed that one of the statements is true and one is false. That is, the task has changed from yes/no for each of A and B to forced-choice involving both A and B. Now, how confident would you be that A is true? That B is true? How confident *should* you be? Assuming that confidence in A and confidence in B at the yes/no stage (c[A] and c[B], respectively) are independent, the normative level of confidence in A at the forced-choice stage (c[A,B]) is the following:

$$c(A,B) = c(A)[1 - c(B)]/\{c(A)[1 - c(B)] + c(B)[1 - c(A)]\}$$

Using the above example, confidence in A should increase from 80% to 86% and confidence in B should decrease from 40% to 14%.

McKenzie et al. (2001) tested the normative model and three descriptive models by asking participants to report their confidence in general knowledge statements presented individually, and then again when they were arranged in pairs. Participants were told that one statement in each pair was true and one was false. Results from two experiments showed that the normative model did not fare well. In particular, participants' forced-choice confidence reports were not extreme enough; they tended to fall somewhere between the normative forced-choice response and their reported yes/no responses. A descriptive model indicating that confidence in the non-focal statement was underweighted tended to perform best. That is, confidence in B did not have enough impact when reporting confidence in A at the forced-choice stage (McKenzie 1998; 1999).

McKenzie et al. (2000) pursued a possible explanation of why confidence in the non-focal alternative was underweighted: Maybe participants did not fully believe that the two statements at the forced-choice stage were mutually exclusive and exhaustive, despite being told so. Indeed, participants are sometimes told that options are mutually exclusive and exhaustive when in fact they are not (e.g., Glanzer & Bowles 1976; Wixted 1992). The authors derived a new normative model (the "trust model") that does not assume that the forced-choice items are believed to be mutually exclusive and exhaustive. The only free parameter in the trust model is the participants' perceived reliability, *r,* of the experimenter:

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

$$c(A,B) = \frac{rc(A)[1 - c(B)] + (1 - r)c(A)c(B)}{r\{c(A)[1 - c(B)] + c(B)[1 - c(A)]\} + (1 - r)\{c(A)c(B) + [1 - c(A)][1 - c(B)]\}}$$

The equation provides the Bayes-optimal forced-choice confidence in A, given r, which is assumed to range between 0.5 and 1. When $r = 0.5$, experimenter reliability is at a minimum, and $c(A,B) = c(A)$. This is as it should be because the experimenter's claim that the forced-choice options are mutually exclusive and exhaustive is seen as completely uninformative and hence confidence in A remains unchanged. When $r = 1$, the experimenter is believed completely, and it is easy to see that the trust model reduces to the normative model provided earlier. When $0.5 < r < 1$, the experimenter is seen as somewhat reliable, and confidence will fall somewhere between c(A) and the normative response provided earlier.

In an experiment using a visual identification task, McKenzie et al. (2000) found that, among the four models tested (all with one free parameter), the trust model performed best (with its free parameter, r, equal to 0.85, indicating that participants as a group did not find the experimenter perfectly reliable). In addition, when participants were asked at the end of the experiment whether they believed that exactly one statement was true when reporting confidence at the forced-choice stage, about 40% expressed at least some doubt. We found this percentage surprisingly high, especially given that we had taken extra care to ensure that participants believed the information. (Participants were not allowed to proceed to the forced-choice stage unless they verbally expressed to the experimenter that they understood that exactly one statement in each pair was true.) Furthermore, when the trust model was fit separately to subgroups of participants: those who did versus those who did not report fully believing the information, r was much higher for the former subgroup (0.94 vs. 0.80). The second experiment manipulated participants' degree of belief that the forced choice items were mutually exclusive and exhaustive and found similar results. Thus, the fact that the traditional normative model did not fare well in the original experiments reported by McKenzie et al. (2001) can be almost entirely accounted for by the fact that at least some participants did not fully believe that the forced-choice items were mutually exclusive and exhaustive, despite being told so.

The trust model's larger implications may be considerable. The idea that participants might not fully believe experimenters' assertions about task parameters is not one that is usually taken into consideration in experiments designed to assess whether participants behave in a normative manner. Often, such studies lead to the conclusion that participants behave non-normatively, just as those of McKenzie et al. (2001) did. For example, Kahneman and Tversky (1973) argued, using results from their well-known lawyer-engineer problem, that participants were committing a normative error by severely underweighting base rates when reporting subjective probabilities. However, Gigerenzer et al. (1988) disputed that conclusion because a key assumption underlying the normative model – that the observations on which the subjective probabilities were based were randomly sampled – was merely asserted by the experimenters (and was untrue). If participants did not believe the assertion, then calling the subsequent responses "errors" would be misleading. Gigerenzer et al. made random sampling more believable to participants by having them draw the observations themselves from urns, which resulted in subjective probabilities closer to the normative response. Gigerenzer et al.'s approach to participant skepticism was to try to eliminate it. However, the results of McKenzie et al. (2000) indicate that this is not sufficient. Some participants will still be skeptical. (Again, the skepticism would be justified: Gigerenzer et al.'s "random sampling" was illusory.) McKenzie et al. (2000) took the additional steps of deriving and testing a normative model that took into account participants' degree of belief in an important task parameter. Note that, in theory, the same could be done for the base rate task.

Regardless of whether or not one agrees with H&O's prescriptions regarding the use of deception, it should be kept in mind that, even if deception were completely eliminated today, participants' skepticism would continue. Such skepticism can be minimized (the first author of this commentary has considered putting a "guarantee" on his laboratory wall, with a picture of his smiling face and the claim, "No deception or receive double your experimental credit or pay"), but it probably cannot be erased completely. Accepting that participant skepticism is an important (and tractable) variable in laboratory experiments, especially those that compare behavior to a normative standard, will lead to more accurate – and interesting – accounts of human behavior.

Theorize it both ways?

Tim Rakow

Department of Psychology, University of Essex, Colchester, Essex, CO4 3SQ, United Kingdom. **timrakow@essex.ac.uk**

Abstract: Psychologists' lack of methodological uniformity reflects their greater breadth of enquiry than experimental economists. The need for a theoretical understanding of one-shot decisions validates research undertaken without the repetition of trials. Theories tested only with financial incentives may not reliably predict some classes of decision such as those involving health. Undue emphasis on the importance of replication risks the proliferation of theories with limited generalizability.

Hertwig and Ortmann (H&O) quite reasonably point to the diversity of purpose among psychologists as one explanation for the range of methodological practices among psychologists investigating decision making. Their critique that this provides no justification for a failure to consider the impact of variation in experimental practice seems similarly reasoned. However, this broader range of reasons that psychologists have for conducting experiments can be seen to lie behind some of the potential strengths in their approach that are not highlighted in the target article.

With respect to the repetition of trials, it is easy to see how the economists' primary interest in equilibrium behavior results in their fairly uniform experimental practice. However, psychologists need not be ashamed that they also have an interest in one-shot, first-shot, or not-very-many-shots decisions. In fact, not least among the good reasons for determining the impact of feedback is that there are many situations where feedback is limited and decisions are taken long before some equilibrium point is, or can be, reached. Our courts operate with jurors who have not previously served in that capacity, and our hospitals are staffed with young doctors who cannot always "familiarise themselves with all the wrinkles of the unusual situation" in which they are placed. Economists may have grounds for claiming that many counter examples to economic theory arise out of a lack of feedback in experimental studies. However, there is surely a place for a theoretical appreciation of non-equilibrium decisions. Thus, while H&O put forward a methodological challenge to examine the impact of the repetition of trials, there is a corresponding theoretical challenge to "theorize it both ways."

With respect to financial incentives, economists' uniformity in experimental practice follows naturally from their primary interest in maximising wealth, profit or utility. Psychologists' more diverse interests, which further include the structure and process of judgements and decisions (Svenson 1996), may well mean that they are prone to lose sight of performance when these other analytical components are to the fore. This shifting focus may be an additional reason for the diversity in practice noted by H&O. Without doubting the value in understanding how incentives affect performance, again psychologists need have no shame that their interests extend beyond what choices are made, to how and why these decisions are taken. Further to this, it would be helpful to examine whether financial incentives are always a suitable

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

"marker" or "metric" for the utility of choices. For instance, economic theories of utility maximisation are frequently applied to health care decisions. However, people readily recognise scenarios involving life expectancy and money as distinct classes of decision, and how they categorize decisions is seen to be related to their preferences (Chapman & Johnson 1995). The parameters of decisions involving health (such as temporal discount rates) can be quite different from those involving money (Chapman & Elstein 1995). Furthermore, contrary to predictions that might be made on the basis of experiments with financial incentives, people can be reluctant to trade or gamble life expectancy for improved quality of life (Stiggelbout et al. 1996). Thus, there is the possibility that an understanding of some classes of decision are best served by experiments involving non-financial incentives.

H&O express the concern that methodological diversity makes replication harder. However, if experimental economists' uniformity of practice improves their chances of replicating results, this may be at the expense of the generalizability of theories. Restricting the variation in experimental practices restricts the range of conditions under which a theory is tested. Of course, this may often reflect that the theory was conceived under strict assumptions. Nonetheless, this exposes the experimenters' endeavors to a number of threats to construct validity such as mono-operation bias, mono-method bias, or restricted generalizability across constructs (Cook & Campbell 1979). The danger is that theoretical constructs are "underrepresented," limiting the application of the theory to a narrow set of situations. The failure to replicate findings in different settings can reflect that a theory has "bitten off more than it can chew" – but that is a lesson we can learn from. Conversely, if in our desire to replicate results we limit the breadth of our enquiry, we are in danger of developing malnourished theories which permit only slow progress.

In summary, there is a sound theoretical rationale for diversity in experimental practices. Namely that diversity in research practice supports the exploration of a range of phenomena under different conditions, without which theories are in danger of having limited coverage or application. To this end, the systematic variation of experimental methodology (as proposed by H&O) seems far preferable to either the restriction of methodological variation or to non-systematic experimental variation.

The game-theoretic innocence of experimental behavioral psychology

Don Ross

School of Economics, University of Cape Town, Rondebosch 7700, South Africa. **dross@humanities.uct.ac.za**
www.commerce.uct.ac.za/economics

Abstract: Hertwig and Ortmann imply that failure of many behavioral psychologists to observe several central methodological principles of experimental economics derives mainly from differences in disciplinary culture. I suggest that there are deeper philosophical causes, based (ironically) on a legacy of methodological individualism in psychology from which economists have substantially cured themselves through use of game theory. Psychologists often misidentify their objects of study by trying to wrench subjects out of their normal behavioral contexts in games.

An economist reads Hertwig and Ortmann's (H&O's) report on the methodological practices of behavioral psychologists with initial incredulity. It is difficult to imagine scientists doing behavioral experiments and hoping for both reliability and replicability without scripts or familiarization regimes or clear incentives for subjects. This astonishment likely depends, however, on an assumption that these psychologists are studying the same sorts of problems we are, and with the same explanatory ends in view. I have no general quarrel with H&O's analysis or their conclusion. They are substantially correct that because experimental economists, unlike psychologists, have not been able to take the appropriateness of behavioral experimentation for granted within their own disciplinary milieu, they have had to be more self-conscious about their methods. Nevertheless, there are at least two broad issues here, one epistemological and one related to policy-related motivations for research, that the authors do not directly address.

H&O rightly note that "the appropriateness of a design depends crucially on what aspects of behavior and cognition a given theory is designed to capture." They claim, also correctly, that "many . . . psychological judgment and decision-making theories are not explicit about the kind of behavior they target – first impressions, learning, or equilibrium behavior – and also do not explicate how feedback and learning may affect it." If psychologists tend to be less *ex*plicit than economists about these things, though, perhaps we might be charitable and ask about the extent to which their general epistemological targets and assumptions are taken to be reasonably *im*plicitly clear – and implicitly clearly different from those of a typical economist.

Cognitive psychologists work mainly within a framework that assumes a "normal" baseline of competences. This norm may be biological and pre-social; that is, they are sometimes wondering about the functional characteristics of the natural, genetic equipment with which people confront the world prior to acquisition of the repertoire of problem-solving routines encapsulated in their particular cultures. On other occasions, they presuppose a maturation threshold that may incorporate some sort of generic cultural input, but imagines leaving the task of specific local cohort contributions to sociologists and anthropologists. Now, to the extent that these assumptions guide psychological practice, that practice may be subject to the critique of Tooby and Cosmides (1992) that it implicitly posits overly general learning mechanisms, with respect to which the relative contributions of nature and nurture – or, more precisely here, individual competence and cultural competence – can be factored out. But this critique is distinct from H&O's. An interest in "baseline" individual competence, encouraged by the traditional basis for distinguishing between psychology and sociology, can explain, even if it does not in the end justify, reluctance to worry about learned behavior at equilibria that may be path-dependent within particular cultures. This point applies in a fairly obvious way to the absence of familiarization regimes in psychological experiments noted by the authors, but it is also relevant to the relative paucity of scripts. The roles around which scripts are cast may presuppose conventional selections of equilibria and their associated strategy mixes in populations, whereas the psychologist may take herself to be seeking the "natural" strategies that subjects tend to use just when the signals and cues that pick out their local equilibrium strategies are missing. Again, standard psychological methods might not be well adapted to finding these strategies; the point is merely that failure to use scripts and protocol familiarization is not inexplicable, and need not be a product of mere sloppiness, disciplinary inertia, or some other aspect of mere disciplinary culture.

This epistemological point is closely related to one concerning policy-related motivations. Economists are typically concerned with promoting predictions of responses to shifts in incentive landscapes. To this extent, they want to know how a particular population, given the conjunction of its members' genetic competences and their conventional equilibrium strategies, will react to exogenous changes in incentivizing parameters. A psychologist is more likely to be hoping to find, and perhaps measure, extents to which individual biases will produce statistical departures from the predictions of rational-agent (both folk and scientific) frameworks. This may explain the relative infrequency of use of monetary incentives in psychology, since these will tend automatically to pull subjects into strategic mode, and this may be exactly what the psychologist's interests imply discouraging.

These points may well deepen H&O's critique, implying a confused fundamental epistemology and metaphysics of behavior on psychologists' part in addition to a shallower lack of methodolog-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ical self-consciousness. For people are no more "natural," as agents, when they are confused by the absence of equilibrating cues (including, then, social roles as represented in scripts, and social learning, and price signals), than they are in a "natural" ethological state when they are naked. Many psychologists may be using (roughly) the *right* methodology to study a kind of phenomenon – the human agent wrenched out of any well-defined game – that naturally occurs only under unusual circumstances in which all decision-problems are purely parametric. Of course, this very kind of strange situation – Robinson Crusoe alone on his island – was just where economists came in to begin with, and what game theory has been moving them away from over the past few years. Perhaps we should not be surprised if psychologists are trailing a bit behind in this respect. If I am right, however, there is an interesting irony here. Economists are regularly criticized for carrying methodological ("metaphysical" would be the more appropriate word in the typical context of debate) individualism to silly extremes. The game-theoretic perspective, however, is a deep cure for this, since its fundamental objects of study are *sets* of behavioral strategies, and individual intentions and competences slip inside black boxes. I am not sure that H&O quite intend to be recommending that psychologists shake off their individualism and get with the game theory program; but I have tried to suggest that they should be read that way, and that so read they are right.

Form and function in experimental design

Alvin E. Roth

Department of Economics and Harvard Business School, Harvard University, Cambridge, MA 02138. **Al_roth@harvard.edu**
www.economics.harvard.edu/~aroth/alroth.html

Abstract: Standard practices in experimental economics arise for different reasons. The "no deception" rule comes from a cost-benefit tradeoff; other practices have to do with the uses to which economists put experiments. Because experiments are part of scientific conversations that mostly go on within disciplines, differences in standard practices between disciplines are likely to persist.

As an experimental economist who finds much to admire in experimental psychology (and has sometimes published in psychology journals and experienced the different expectations of referees and editors in the two disciplines), I read Hertwig and Ortmann (H&O) with lively interest. Although I am not sure that the use of "scripts" is common in economics experiments, I can confirm from experience (e.g., as coeditor of the *Handbook of experimental economics,* Kagel & Roth 1995) that economists do indeed almost universally use performance-based monetary incentives, frequently examine repeated trials with feedback, and almost never use deception. I recognize the benefits to experimental economics that H&O associate with some of these experimental methods. I also think both economics and psychology reap considerable benefits by allowing experimenters flexibility. In experimental design, form follows function, and different design practices reflect some of the different uses to which experiments are put. In this connection, I think economists might also profit by making more use of some practices from psychology.

Let me start with an anecdote about the different perceptions of deception. When I was setting up a new lab, I met with a research administrator, a psychologist. One of my colleagues mentioned that economists do not use deception, and the administrator looked to me, in puzzlement. I said economists believed a reputation for non-deceptive practices was a public good; that if we started to use deception, this might cause changes in subject behavior and make experiments harder to interpret. The administrator replied, with apparent relief, "I knew economists wouldn't object to deception on ethical grounds."

This practical concern may have different implications for psychologists and economists, given that economists do not have a long history of interesting uses of deception in experiments, while psychologists do. It is difficult to get large, diverse groups to provide public goods, especially once there is a history of free riding. Particularly for the public good "no deception," even if all psychologists stopped using deception tomorrow, the fact that important experiments using deception are taught in introductory classes might mean the benefit from this change would be long in coming, since psychology students would remain suspicious for a long time. But the costs would be immediate.

And there are costs, because there have been psychology experiments that used deception to spectacularly good effect. For example, the widely taught Milgram experiment would lose its force if some subjects had not believed they were administering dangerous shocks to an unwilling subject. (And even economists might find ethical objections to the non-deceptive version of the experiment in which they were.)

For economists, in contrast, the fact that deception is seldom used and is received with hostility by referees and editors makes the non-deceptive status quo easy to maintain. I tell students they shouldn't contemplate using deception unless they have a spectacularly good reason. The tradeoff of private costs and public benefits may thus be different for economists than for psychologists.

The other differences in experimental methods between disciplines often have to do with the different ways experiments, and theories, can be used. The traditional use of experiments is to test if theories are false, and a theory can be falsified with a single well-chosen experiment. However, if the theory is intended to be a useful approximation, valued for the accuracy of its predictions on many, but not necessarily all domains, then the importance of a falsifying experiment may be ambiguous. Economic theories are mostly meant to be useful approximations, and many of the experimental practices in economics arise from the need to persuade other economists that the experiment is appropriate for evaluating the usefulness of the approximation.

The choice of single or repeated trials is often made with this in mind. For a task in which behavior changes substantially over repeated trials with feedback, implications for behavior outside the laboratory will depend on judgments about the natural environment in which similar decisions may be taken. Are we trying to predict the behavior of people with lots of experience at the task, or of people who have little relevant experience? Which kinds of experience are relevant? Economists (like psychologists) are interested in behavior both in novel and in familiar situations.

Because economists are often interested in applying very general theories over a wide domain, economic experiments are often conducted in relatively abstract, context-free environments. If the experiment tests a theory that predicts that players in a game tend to play an equilibrium because they are highly rational, the best evidence might come from an experiment in which there are no cues to the subjects about how to behave except those strategic features of the game that the theory predicts are important. In a high context experiment, subjects might perform as predicted for different reasons. But while abstract environments may be good for testing the most general predictions of the theory, behavior in abstract environments may not always be a good predictor of behavior in familiar environments. In this respect there is room for economists to emulate psychologists by more frequently framing experiments in natural contexts.

Monetary incentives started to to be fairly standard in economics experiments as early as the 1950s, to make clear the economic structure of the experimental environment (cf. Kagel & Roth 1995; or Roth 1993). Even here, relying exclusively on purely monetary incentives would miss some important phenomena. Kahneman and Thaler, for example, showed that subjects who were given a coffee mug valued it more highly than subjects who had not been given one. To the extent that people treat possessions differently from money, this would have been a hard effect to observe if the only payoffs available to subjects had been monetary. On the other hand, even when social incentives are the ob-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ject of study, their importance is often clarified (at least to economists) when they can be seen to overcome contrary monetary incentives.

In summary, a good experiment is one that studies an interesting question in a way that controls for the most plausible alternative hypotheses. Since "interesting" and "plausible" are in the eye of the beholder, we can expect persistent differences in good experiments across disciplines, despite the fact that we nevertheless have a lot to learn from one another.

From old issues to new directions in experimental psychology and economics

Vernon L. Smith

Economic Science Laboratory, University of Arizona, Tucson, AZ 85721.
smith@econlab.arizona.edu

Abstract: The rhetoric of hypothesis testing implies that game theory is not testable if a negative result is blamed on any auxiliary hypothesis such as "rewards are inadequate." This is because either the theory is not falsifiable (since a larger payoff can be imagined, one can always conclude that payoffs were inadequate) or it has no predictive content (the appropriate payoff cannot be prespecified).

Scripts and rewards in the early years. From 1955 to 1965 certain founders of experimental economics included both psychologists (W. Edwards, S. Siegel, A. Tversky) and economists (R. Selten, L. Fouraker, V. Smith, J. Friedman). These researchers relied on monetary rewards for motivation; some varied the rewards as treatments; and they all used careful scripts to limit task ambiguity. The excellent instructions and rigorous protocols pioneered by Siegel set protocol standards – even much of the language – that were later applied widely by experimental economists, but not psychologists. Why the two communities have diverged is an important topic in experimental methodology that is explored in the interesting essay by Lopes (1991), while the new understanding contributed by Gigerenzer may hasten convergence. Psychologists must seek a deeper understanding of the observational phenomena to which they attach names, such as framing, representativeness, anchoring, availability, and fairness.

Games theory, experiment, and the Duhem-Quine (D-Q) problem. At the heart of the Hertwig & Ortmann (H&O) target article is the D-Q thesis: all tests of theory require auxiliary hypotheses to implement them. Consequently, all falsifying outcomes can be dismissed by arguing that it is the auxiliary hypotheses, not the theory, that must be rejected. The faith of devout believers stands unshaken, while the skeptics say, "We told you so." The former, after seeing the outcome, seek a reinterpretation of what constitutes an appropriate test; the latter seek new falsifying outcomes. Thus, game theory often fails, based on the classic assumptions of dominant strategy self-interested types, common "knowledge" of the same, and backward induction. Here are a few examples of *ex post facto* reinterpretations when tests of game theory fail.

Treatment protocol and/or context
Face-to face bargaining
Anonymous dictator, ultimatum, and other extensive form games
Same; any experiment
Same, but use advanced graduate students
Any experiment

Theory rescuing interpretation
Social context loses control over preferences
Reparameterize with other-regarding utility to fit the data
Unsophisticated undergraduates
All subjects need experience: repeat with new strangers after each play
Payoffs too low

These are the "just so stories" of economic/psychological testing. Taking seriously the last, most ubiquitous objection to falsifying evidence, there are only two interpretations: (1) the theory cannot be falsified; (2) the theory has no predictive content. (1) follows if, for any increase in payoffs, the test is negative; then one can always imagine still larger payoffs, and argue that the payoffs were inadequate. (2) follows if, after increasing payoffs sufficiently, the test is positive; then the theory is sterile because it cannot prespecify the payoff conditions that enable successful testing. The same logic applies if a falsifying outcome is rejected because the subjects were insufficiently sophisticated. The theory forever lags behind the empirical results, yielding what Lakatos calls "miserable degenerating research programmes."

This undesirable state is a consequence of the rhetorical commitment to falsificationist/predictive criteria. Why should we believe that we can construct falsifiable, predictive models by abstract thought alone? If we have learned anything in 50 years of experimental economics it is that real people do not solve strategic decision problems by thinking about them the way we do. In fact, we do not solve our own decision problems this way, except in our publications. There isn't time, it's cognitively too costly; and if the problem has really high stakes (the vast majority of daily decisions have low stakes), we hire professionals.

All science lives in a sea of contradictory observations. Our task should be to modify theory in the light of evidence, and aspire to encompass suspected auxiliary hypotheses (stakes, subject sophistication) explicitly into the theory to motivate new tests. If payoffs matter, why? All theories "predict" optimal choice, however gently curved the payoff function. If results improve with reward some cost must be overcome by the increased motivation. Walker and Smith (1993) assumed that cognitive effort cost, and the productivity of effort, lay behind the data, but that such "unobservables" had observable consequences. (See Smith & Szidarovszky 2000 for an extension to strategic interactions.) More constructivity is needed rather than to repeat over and over that payoffs are inadequate. The latter leaves unanswered why "small" payoffs sometimes yield good results, and why increasing payoffs do not always yield better results.

New directions defined by the machine builders. Science is driven more fundamentally by the machine builders, than either the theorists or experimentalists. Witness the impact of the telescope, microscope, and accelerator in changing the terms on which theory interacted with observations. Computer/communication and imaging technologies are driving experimental economics in new directions. Both will marginalize extant research in individual decision.

When A. Williams programmed the first electronic double auction (e-commerce in the lab) in 1976, it changed the way we thought about markets, much as the internet is changing the way people think about doing business. Circa 1976 we thought going electronic would merely facilitate experimental control, data collection, and record keeping. What we discovered was altogether different: computerization vastly expanded the message space within which economic agents could communicate at vanishingly small transactions cost. This enabled Stephen Rassenti to invent the first smart computer assisted combinatorial auction market, driven by optimizing algorithms applied to the decentralized messages of agents (see Rassenti 1981). This set the stage for smart markets in electric power, pipeline networks, water networks, scheduling space manifests, and pollution credits in which commodities are endogenously defined by a market for their characteristics. Lab experiments became the means by which heretofore unimaginable market designs could be performance tested.

Brain scanning technology now enables us to study the neural correlates of "mind reading" (inferring the thoughts of others from their words or actions) and, for the first time, associate mental modules in the brain with external choices. As this technology becomes cheaper and less cumbersome to use, it promises to bring new dimensions of understanding to our traditional observations from experiments. Game theory postulates known intentions, and

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

minds do not have to be read. But when subjects forego the subgame perfect equilibrium outcome to attempt cooperation, and this trust is reciprocated, it is interpreted as a mutually beneficial exchange based on "mindreading." McCabe et al. (2000) report the first evidence that such decisions activate brain modules that are associated with mentalizing about what players believe about each other.

Different perspectives of human behavior entail different experimental practices

Ramzi Suleiman

Department of Psychology, University of Haifa, Haifa, 31905, Israel.
suleiman@psy.haifa.ac.il

Abstract: My main argument is that the advice offered to experimental psychologists by Hertwig & Ortmann overlooks fundamental differences between the goals of researchers in psychology and economics. Furthermore, it is argued that the reduction of data variability is not always an end to be sought by psychologists. Variability that originates in individual differences constitutes valuable data for psychological research.

Hertwig and Ortmann (H&O) discuss four key features of experimental design (enactment of scripts, repetition of trials, performance-based monetary payments, and the use of deception) that are realized differently in economics and in areas of psychology relevant to both economists and psychologists. Notwithstanding some minor reservations, the authors express a strong preference for economists' practices on all four features and advise experimental psychologists to adopt the conventional practices of experimental economics. They also posit that the arbitrariness of experimental designs in psychology is partly a byproduct of theories which lack explicitness about the "kind of behavior they target"; this in contrast to the clarity of (the dominant) game theory in economics which specifies equilibrium behavior as its target.

I would like to suggest that the advice offered to experimental psychologists by H&O is highly sweeping and that it overlooks fundamental differences between the theoretical perspectives and goals adhered to by researchers in the two disciplines. A detailed substantiation of my claim is beyond the limited scope of this commentary, thus I shall focus on two critical remarks.

(1) H&O applaud the implementation of clear (hence binding) scripts in experiments conducted by economists. They argue that scripts can constrain participants' interpretations of the situation by focusing their attention on those cues that are intentionally communicated by the experimenter (e.g., the task instructions), thus clarifying the demand characteristics of the social situation. At least two pitfalls of using scripts that are too restricting may be pointed out. The first is methodological and is relevant to experiments in psychology and economics alike. It concerns the undesirable possibility that "clarification of the demand characteristics" of the type described above is unavoidably entangled with the enhancement of those demand characteristic which coincide with the experimenters' focus. A relevant argument was succinctly put by Dawes (1999), who posited that "the psychologist critic concerned with experimental demand would point out that economists' experiments that are carefully and tightly designed to extricate rational behavior involve 'some beating subjects over the head' through constant repetition, feedback, complete and detailed information, and anonymity" (Dawes 1999, p. 23).

The second argument against the use in psychological research of scripts and other design features that are too "tight" rests on the fundamental difference between theories in psychology and economics with regard to what determines human behavior. Economic theories, including game theory, are hardly concerned with personality related variables. Such theories and related analyses of experimental data assume that behavior is determined by situational factors (e.g., the game structure). In contrast, psychologists are interested not only in situational effects on behavior, but also in the effects of personality variables and their interaction with a given situation. Such interactionist perspective (Blass 1984) is fundamental to the social psychological paradigm (Lewin 1935; 1947). From this perspective, design features that are too "tight," such as scripts that are too constraining and repetition of trials, are bound to enhance the salience of situational factors at the expense of personality (individual differences) related factors. This will eventually impede the detection of possible effects of individual traits and styles on behavior.

The above argument relates to the strong emphasis put by H&O on reducing data variability as a desired objective. No one can argue against the importance of controlling for intervening variables and the minimization of measurement noise. Nonetheless, mere reduction of data variability is not an end to be sought by psychologists. For them, data variability which originates in individual differences (in response to a given situation) is a main ingredient of the "bread and butter" of their research.

To summarize this point, it is argued that the stand taken by H&O on this issue is too simplistic. Obviously, designs that are free from constraint on subjects' choices are useless because they do not yield interpretable information. But, conversely, under total constraints "no information can be gathered either because subjects 'could not have done otherwise'" (Dawes 1999, p. 22). Rather than unconditional and categorical choice between options, it is proposed here to view the extreme cases described above as two endpoints on a continuum, and to subordinate the choice regarding the desirable degree of constraint to the specific research objectives.

(2) The authors posit that if theories in psychology were more explicit about their target behavior (as game theory is), then the theories rather than the experimenter would define the appropriate test conditions. Notwithstanding the value of theory (preferably competing theories) in guiding empirical research, it is argued that the economic research on equilibrium behavior is not a good example to follow. First, because it devotes little effort to "bottom up" investigations aimed at using empirical results to modify existing theories and stimulate novel development of theory. Second, because, like any theory-centered approach, it is likely to result (as it frequently does) in procedures that are biased toward obtaining a theory-predicted result (Greenwald et al. 1986). A demonstrative (albeit extreme) example of such theory-confirmation bias is evident in the research on testing the equilibrium prediction in ultimatum bargaining. In one study (Binmore et al. 1985) the written instructions included the following passage: "How do we want you to play? YOU WILL BE DOING US A FAVOR IF YOU SIMPLY SET OUT TO MAXIMIZE YOUR WINNINGS" (capitals in the original). Other studies were less explicit about directing subjects to behave strategically, nonetheless they included various extrinsic manipulations, designed to elicit strategic behavior from the side of allocators. Auctioning the allocator's role (Güth & Tietz 1986), enforcing the allocator's property rights, and "framing" the situation in economic terms of "buyer" and "seller" (Hoffman et al. 1994), are all examples where intervening variables that are extrinsic to the game (and thus not alluded to by game theory), were introduced with the purpose of "pulling" subjects' responses towards validating the theory where the basic game fails to do that.

Self-interest as self-fulfilling prophecy

Mark Van Vugt

Department of Psychology, University of Southampton, Southampton, United Kingdom. **mvv@soton.ac.uk** **psy.soton.ac.uk/~psyweb/vugt**

Abstract: The adoption of experimental methods from economics, in particular script-enactment, performance-related payment, and the absence of deception, will turn experimental social psychology into a trivial science

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

subject. Such procedures force participants to conform to a normative expectation that they must behave rationally and in accordance with their self-interest. The self-fulfilling prophecy inherent in these procedures makes it more difficult to conduct innovative social-psychological research.

Why are the experimental procedures in psychology so diverse compared to the highly standardized method in experimental economics? The obvious answer is that it reflects a difference in theoretical diversity. Whereas experimental economists use the theory of self-interest as a unique explanatory framework for understanding human behavior, psychologists have no single motivational theory of human conduct. Psychologists believe that, given the right conditions, people can be rational or irrational, selfish or altruistic, aggressive or helpful.

Largely ignoring these disciplinary differences, Hertwig & Ortmann (H&O) propose that experimental psychology could benefit from adopting the procedural regularity of experimental economics with the aim to "reduce participants' uncertainty by precisely specifying the social situation." Inspired by the research tradition in economics, they suggest four ways to improve the experimental practice in psychology, via script enactment, repeated trials, performance-related payment, and the absence of deception.

Although I share some of H&O's concerns about the experimental and mundane realism of psychological research, I believe the introduction of these practices has potential drawbacks, in particular for experimental social psychology, my research field. Because the laboratory experiment is an artificial, isolated event without the normal situational constraints on behavior, participants are extremely sensitive to cues from the experimenter. Interestingly, this is even more so when people are keen to participate, for example, because they receive a monetary reward, than when they participate in order to receive course credits (Rosenthal & Rosnow 1975).

Consequently, participants will respond to the slightest suggestion about the appropriateness of certain attitudes or actions. The result is that participants' behavior simply becomes a self-fulfilling prophecy, a confirmation of the experimenter's pre-existing belief set (Snyder & Swann 1978). Unfortunately, this obscures the influence of many interesting psychological differences between people, such as differences in personality, attitude, and values.

First, the presentation of script information before the experiment can act as a self-fulfilling prophecy, because scripts are seldom value-neutral. For example, asking participants to imagine that they are traders in an investment game elicits their competitive attitudes and beliefs, hence their desire to win the game. Asking the same person to act as a social worker or teacher increases their cooperation and trust. Participants simply respond to the script instead of the nature of the task. In social-psychological research it is therefore advisable to give as little information about the roles of participants, unless it is of theoretical interest (e.g., leader-follower).

The danger of a self-fulfilling prophecy effect is even more apparent in tasks which use performance-related payment. I am not convinced about the argument put forward in the target article that experiments with performance payment increase the motivation of participants. A plausible alternative explanation for the apparent superior performance in tasks with payment schemes is that it sustains a norm of self-interest (Miller 1999). Participants believe they should behave in a particular way, because they are paid for their performance. No wonder then that participants' responses show little deviation from the rationality principle. They simply behave according to what they believe the experimenter is looking for, that is, to act in their self-interest.

This is even problematic in experiments where performance is not defined in terms of individual gains but in gains for others, for example in the study of prosocial behavior. If research shows a greater incidence of helping another person in reward-contingency tasks, what can be concluded about the motives for helping? Either participants use the reward as a legitimate reason for helping (in which case they are selfish) or they use it as a post hoc justification. We will never know.

With regard to the ethics of psychological research, I agree that the use of deception is a questionable practice and that it should be avoided as much as possible. However, deception is permitted, I believe, if it serves the goal of developing theories of human behavior which are counter-intuitive. For this purpose, sometimes researchers must conceal the true purpose of an experimental task.

One of the more robust effects in social psychology, the cognitive dissonance effect (Festinger 1957), could have only been found because participants were giving misleading reasons to engage in actions countering their initial attitude. For example, had participants in advance been told that the true purpose of a study was to find out how they would change their attitude, for example, towards abortion, as a result of writing a pro- or anti-abortion essay, they would not have shown the dissonance-effect (i.e., change in initial attitude). They would have stayed with their initial attitude, believing that this was the rational, hence, more rewarding choice. Transparency of instructions thus carries the danger of leading participants to conform to certain norms and expectations outside the experimental situation.

Interestingly, the practice of deception is attacked on external rather than internal grounds in the article. Deception does not so much threaten the internal validity of research, but it erodes the general faith in experiments. Participants will be less likely to turn up again. Although I endorse this view (and believe therefore in an extensive debriefing procedure), the same argument can be made against the use of payments. According to economic simple reasoning, once participants have earned a particular sum of money in an experiment, they will be less motivated to do their best in experiments which offer less. This could lead to an inflation in payment fees, which no department, economics or psychology, is able to afford.

In sum, I have argued against the application of methods from experimental economics to experimental social psychology: in particular, script-enactment, performance-related payment, and absence of deception. These practices turn social psychology into a trivial science. They force participants to conform to the normative expectation to behave rationally and in line with self-interest. The self-fulfilling prophecy inherent in such procedures hinders the scientific progress by decreasing the chances of finding surprising and counter-intuitive results.

Meta-theory rather than method fascism

Elke U. Weber

Graduate School of Business and Department of Psychology, Columbia University, New York, NY 10027-6902. **euw2@columbia.edu**

Abstract: Three comments take issue with specifics of the target article. First, I argue for the development of meta-theory to deal with inconsistencies in experimental results. Second, I advocate the use of the compatibility between experiment and application to decide on the optimal design and procedure of experimental studies. Last, I support the explicit incorporation of motivation into models of decision making.

Hertwig & Ortmann (H&O) are to be commended for their comparison of the methodological practices of psychology and experimental economics with the goal of making the behavioral decision research enterprise more cumulative. While globally sympathetic and locally mostly in agreement, I have three comments that qualify or take issue with some specifics of their article.

H&O explicitly disavow the rather rigid methodological prescriptions of experimental economics (the "methodological fas-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

cism" facetiously alluded to in the title of my commentary). Yet, for reasons that are sound and well articulated, their implicit sympathies lie closer to the economics' end of the continuum than to the laissez-faire state of affairs of psychology. I would like to argue against both the restrictive practices of experimental economics and against the authors' final recommendation, the ecumenical "do-it-both-ways" rule, which may be too undirected.

(1) As others and I have argued elsewhere (see Weber & Hsee 1999), empirical research ought to be theory-driven. Existing theory can and should guide questions about both the design and the analysis of experiments. Meta-theory (i.e., "meta" in the sense of taking existing theory one level further back) can also help reconcile apparently contradictory empirical results. Erev et al. (1994) provide a classic example for how the introduction of an additional theoretical construct – the existence of random error in judgments of likelihood – resolves the apparent contradiction between two sets of research findings, namely, conservatism in the revision-of-opinion literature and overconfidence in the calibration literature. I argue that meta-theory can also shed some light on the effect of enacting a script discussed by H&O in section 2. Being in a particular social or economic role often comes with a specific set of asymmetric loss functions. Thus, the buyer of a used car may feel worse if he overestimated the car value and paid too much for it, than if he underestimated the value and offered too little, running the chance of being outbid by other interested parties. A different set of asymmetric loss functions are in place for the seller of the car or for a neutral assessor of the car's value (Birnbaum & Stegner 1979). Asking respondents to provide a value for a commodity of uncertain value without providing them with a script that specifies the perspective they should assume may thus lead to inconsistent and "noisy" answers, because different respondents will (implicitly) assume different perspectives, perhaps as a function of recent prior experience. Asymmetric loss functions affect judgment and choice in a wide range of situations (Weber 1994), leading to decisions that often deviate from expected utility maximization and can be modeled by rank-dependent utility maximization where the nature of the rank-dependency of utilities can be predicted by the nature of the loss function asymmetry (Weber & Kirsner 1997). The perspectives example mentioned by H&O as leading to different problem-solving results as a function of solving it from an employer's or employee's perspective (sect. 2.1), is another case in point. Researchers would thus be well advised to think about the possible perspectives participants could bring to a particular judgment or choice task and to consider the implications of each perspective for the experimental task, starting with a specification of loss function symmetry or asymmetry. Experimental specification of a script that induces a particular perspective will then lead to more than just random error reduction, but will allow us to test hypotheses about the mechanisms by which perspective affects task performance. Meta-theory will also be required to explain the inconsistent effects of financial incentives discussed by the authors in section 4.

(2) Compatibility is a construct used by Slovic et al. (1990) to model the effect of response mode (e.g., pricing vs. choice) on apparent preference (see Payne et al. 1993, pp. 43–45). A more macro-level compatibility principle can be brought to bear on the question of optimal design of experiments or, perhaps more pertinently, on the question of optimal generalization from existing experimental results. Given that response mode, framing manipulations, and script-induced perspectives have been shown to affect judgments and choice, the question arises of what behavior to predict in a particular policy application or how to solicit preference in a particular decision-aiding application. For both questions, the compatibility-principle can provide guidance: Make the experiment as similar as possible to the application on all features of context, perspective, and elicitation procedure that are known to (or suspected of) affecting task performance. Experiment-application compatibility will maximize predictive success in the policy application and minimize any differences between predicted and experienced utility in the decision-aiding application. The implication for the target article is that the study and incorporation of the effects of perspective and role playing into judgment and choice theory and experimental design will allow us to apply the compatibility principle in a more effective fashion. Some applications may best be modeled by putting respondents into a particular well-specified role with the help of a detailed script. Other applications may be better modeled without such a script because decision makers in the application are similarly unsure of the role they play in the situation. The compatibility principle similarly applies to H&O's discussion of the pros and cons of repeated trials versus snapshot studies in section 3. Some applications provide decision makers with feedback, making learning and equilibrium models relevant theoretical tools and repeated trials the appropriate experimental procedure. Other applications do not, and therefore ask for theory and procedure capable of modeling first impressions or untutored choice.

(3) Finally, I would like to second H&O's exhortation to incorporate decision makers' motivations more explicitly into judgment and choice models. In my lab, we are engaged in a research project to show that decision makers' goals and motivations influence how they go about making their decision, and that the chosen decision mode in turn influences the outcome of the decision (Weber 1998). Our preliminary results show the utility of more explicit theory about the effects of motivation on decision processes and outcomes.

Deception by researchers is necessary and not necessarily evil

David J. Weiss

Department of Psychology, California State University, Los Angeles, CA 90032. **dweiss@calstatela.edu**
www.calstatela.edu/academic/psych/html/dweiss.htm

Abstract: Despite claims of pure pragmatism, Hertwig and Ortmann's negative perspective on deception suggests a selfish psychologist willing to sacrifice the reputation of the discipline in order to expedite the research. Although questions that appear to have correct answers may be investigated with complete openness, research that delves into personal secrets often requires deception as a tool to counter self-presentation bias.

From the participant's perspective, there are two kinds of behavioral research. Hertwig and Ortmann (H&O) seem to have considered only the class of experiments in which the task is to perform optimally in some defined respect. I refer to these as "IQ tests." There is clearly a right answer, and the participant is asked to find it. Economists and some psychologists exclusively employ such tasks in their studies. "IQ test" researchers have traditionally seen little need for deception, and so can afford to seize the moral high ground.

The other class of studies consists of those that "pry." The researcher wants to know some aspect of the participant's character, attitudes, or personal history. If we presume that one of the participant's main personal goals in any study is to come off looking good – referred to as self-presentation bias (Catania et al. 1990) – then we can see why the methodologies for the two kinds of research often diverge.

When there appears to be a correct response, the participant can do little to improve self-presentation other than try to get the right answer. On the other hand, when the research appears to be probing personal issues, then many participants inevitably try to hide attitudes or behaviors they think reflect badly on themselves. The researcher's primary tools for countering these attempts to conceal are deception and privacy manipulation. It is important to note that the investigator cannot always specify the sensitive areas (Ong & Weiss 2000).

For me, the danger is that the persuasive arguments raised by H&O will be used by the timorous APA to justify a blanket prohi-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

bition against deceptive practices. The other key variables they mention (scripts, repetition, and performance-based payoffs) seem to concern empirical issues, but deception is a different matter. Despite the authors' explicit statement that they do not oppose deception on moral grounds, I am skeptical. The argument regarding the common good is clearly designed to inflict guilt on those of us who enlist deception in an effort to "pry." Exposed trickery contaminates the pool of potential subjects, so that even the pure suffer for the sins of the miscreants down the hall. I liken H&O to vegetarians who, having failed to convince the carnivores that eating animals is evil, stress the health benefits of the ethical diet.

Continuing in metaphorical mode, I suggest that a superior analogy for deception is the medical community's reliance upon antibiotics to combat bacterial infection. Each time an antibiotic is used, the likelihood of its efficacy in the future is reduced; but curing the current patient is deemed to justify this cost. New antibiotics have to be crafted on an ongoing basis to keep ahead of the microbes; so too behavioral researchers may need to create novel deceptions. The one-way mirror no longer works, so we abandon it. Of course, we always seek superior new methods; we are not committed to deception any more than the physician is committed to antibiotics. We use the most appropriate techniques available.

Even the "IQ testing" research community would be well-advised to be cautious regarding a ban on deception. In psychology, it has become routine to collect demographic information from our volunteers. Some participants may be curious as to why they are asked to provide their ethnicity or gender. That curiosity may lead to hypotheses that will affect the behavior under examination. Some will not want to reveal personal data, and circumventing that reluctance may call for subtlety.

People can be treated with respect even while they are being tricked (Saxe 1991). What is important is that we do not violate our participant's sense of self-worth. That is a subjective construct, to be sure, which is why the democratic institution of the Institutional Review Board can be so helpful.

Individual psychology, market scaffolding, and behavioral tests

Daniel John Zizzo

Department of Economics, Christ Church College, Oxford University, Oxford OX1 3BW, United Kingdom. **daniel.zizzo@economics.ox.ac.uk**
www.economics.ox.ac.uk/research/BREB/index.html

Abstract: Hertwig and Ortmann (H&O) rightly criticize the usage of deception. However, stationary replication may often have no ecological validity. Many economic experiments are not interactive; when they are, there is not much specifically validating H&O's psychological views on script enactment. Incentives in specific market structures may scaffold even zero rational decision-making, but this says very little about individual psychology.

People sometimes ask me what experiments run by economists look like. My short answer is that it is like doing a psychological experiment, but with monetary incentives and no deception. To these factors, H&O add script enactment and the importance attributed to repetition and stationary replication.

I wholeheartedly agree with H&O's treatment of deception. An experimental economist, Bardsley (2000), recently devised a method to reap some benefits of deception without actually deceiving anyone. Take an experiment with repeated tasks and monetary incentives. Subjects are informed in advance that one task is "for real" (they are paid for it, and everything they are told about it is true) and others fictitious, but they are not told which is which. The technique is not without problems. For example, if a subject is surprised by, say, the others' contributions in a social dilemma, he may think it is a fake distribution of contributions, and behave (and not learn) accordingly. Yet, the method is potentially interesting.

I agree less with H&O's stress on stationary replication. How empirically plausible is it that, in many cases, subjects repeat and receive feedback on exactly the same task 10, 20, 100 *n* consecutive times? There are three problems here: the number of times a subject has faced the same task in the real world; the availability of unambiguous feedback; the frequency of presentation. Unless all three criteria are met in the real world environment of interest, there is no reason why limited research resources should be invested in trying to achieve stationarity. My criteria might be too strict because they do not allow for transfer of knowledge between similar tasks. If transfer of knowledge were adequately strong, however, why should it not apply also to when subjects enter the laboratory (Loewenstein 1999)? If anything, strong knowledge transfer effects would make short-run responses more, not less, interesting.

I am not fully convinced about script enactment. First, many experimental economists deal with individual decision-making using non-interactive experiments (e.g., most of the work reviewed in Camerer 1995). Second, economists have an obvious interest in interactive decision-making in specific institutions and games, and have to provide instructions to subjects to deal with this kind of experiment; however, there is very little in this specifically validating H&O's psychological views on script enactment.

When interactive experiments are made, roles are typically assigned only insofar as they are absolutely necessary: this is the case when subjects are not all in the same position, but have to understand what subjects in other positions do; then it is virtually unavoidable to use labels ("buyers," "sellers"). Even then, as admitted by H&O in a footnote, instructions are as content-free as possible. You talk about buyers and sellers, not about workers and firms, even in an experiment on fair wages in sequential labour markets (Fehr et al. 1993). In a trust experiment, you talk about "persons in Room A" and "persons in Room B" – *not* according to the script of a "lost wallet game" used in the published paper (Dufwenberg & Gneezy 2000). H&O believe that scripts matter, and they may be right, but, aside from a footnote, one should not give the impression to ascribe to experimental economists a methodology which is not (by and large) theirs, except where forced by the nature of the object of investigation.

What H&O attribute to script enactment may not have to do with script enactment at all in explaining how bounded rational agents fare in specific market institutions. It is known from computer simulations, for example, that even zero rational traders will converge to the efficient equilibrium in a double auction (Gode & Sunder 1993). This tells us nothing about human psychology. Rather, it tells us about how the structural incentives underlying *specific* institutions help to "scaffold" bounded rational agents in making a more rational choice (see Clark 1997). Similarly, the fact that fairness does not matter in straight double auctions says something about the incentive structure of this peculiar market structure, but nothing about fairness as a psychological motivation; indeed, simply adding a sequential component to the double auction changes the incentives and allows fairness to re-emerge (Fehr & Falk 1999). One cannot use the evidence from market experiments to validate the importance of script enaction and, consequentially, the uselessness of decontextualised experiments. Nor can it be used, say, to dismiss bounded rationality anomalies as figments of semantic misunderstandings.

On this last point, Camerer's (1995) review shows that methodological strictures reduce, but rarely eliminate bounded rationality anomalies (outside, only sometimes, specific market institutions). The conjunction fallacy is a case in point. Zizzo et al. (2000) replicated findings on the conjunction fallacy with monetary incentives, strong hints, the word "likelihood" rather than "probability" and some learning feedback. Recently, I implemented a behavioral version of the conjunction fallacy task, with more salient monetary incentives and 150 rounds of variously composed prac-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

tice; I still found a robust lower bound of 20% (with an average of 36.61%) of fallacy committal (Zizzo 2001). Rather than chasing the goose of trying to wipe out "irrational" behaviour by ever more sophisticated means, researchers should focus on trying to explain *how and why* bounded-rational agents behave the way they do. I am unsure the methodological debate really helps in this respect, though perhaps behavioral analysis as such may help going beyond claims about semantic ambiguities.

Finally, H&O's footnote dismissal of Frey's work on crowding-out of intrinsic incentives is unwarranted. They ignore the field evidence presented by Frey in leading economics journals (Frey et al. 1996; Frey & Obeerholzer-Gee 1997). They also ignore current experimental economics research showing that crowding-out matters in contractual relationships (Bohnet et al. 2000; Fehr & Gachter, forthcoming; Gneezy & Rustichini, forthcoming). Crowding-out of intrinsic incentives should be a concern to psychologists, particularly social psychologists, given the non-financial nature of many non-economic decisions.

Authors' Response

Money, lies, and replicability: On the need for empirically grounded experimental practices and interdisciplinary discourse

Ralph Hertwig[a] and Andreas Ortmann[b]

[a]*Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany;* [b]*Center for Economic Research and Graduate Education, Charles University and Economics Institute, Academy of Sciences of the Czech Republic, 111 21 Prague 1, Czech Republic.*
hertwig@mpib-berlin.mpg.de **andreas.ortmann@cerge.cuni.cz**

Abstract: This response reinforces the major themes of our target article. The impact of key methodological variables should not be taken for granted. Rather, we suggest grounding experimental practices in empirical evidence. If no evidence is available, decisions about design and implementation ought to be subjected to systematic experimentation. In other words, we argue against empirically blind conventions and against methodological choices based on beliefs, habits, or rituals. Our approach will neither inhibit methodological diversity nor constrain experimental creativity. More likely, it will promote both goals.

R1. Introduction

We concluded the target article in the hope that it would "spur psychologists and economists to join in a spirited discussion of the benefits and costs of current experimental practices." We are delighted to see that our hope has become reality. Psychologists and economists, together with researchers from other disciplines, responded to our "gentle aggression" (**Harrison & Rutström**) by contributing to an overdue conversation on the nature, causes, and consequences of the diverging experimental practices in psychology and economics. We would be pleased if this discourse indeed moved us "closer to a common language for scientific discovery" (Harrison & Rutström).

Our reply includes two major parts. In the first part (sects. R3–R6), we address commentators' responses to our discussion of four key variables of experimental design. The second section (sects. R7–R9) is organized around a number of additional issues – among them are conjectures on the causes of the methodological differences and the affinity between experimental economics and behaviorism. We conclude with an outline of some aspects of experimentation in psychology from which we believe economists could learn. We begin with a discussion of what appears to be the most serious and common concern with regard to our analysis.

R2. Do our policy recommendations jeopardize experimental diversity and creativity?

A number of commentators (e.g., and prominently, **Gil-White, Guala, Hilton, Huettel & Lockhead,** and **Kurzban**) argue that our recommendations – to ground design decisions in empirical evidence, to systematically manipulate key variables of experimental design (as expressed by our do-it-both-ways rule[1]), and to use deception as practice of truly last resort – would stifle methodological diversity and constrain experimental creativity. Both are important goods, and endangering them would compromise our policy recommendations. Like **Maratsos** and **Harrison & Rutström**, however, we believe that our recommendations will not have this effect. Rather, we anticipate our recommendations will promote both experimental diversity and creativity. To explain why, we first summarize our argument and – drawing on commentators' objections – refine it.

We documented in the target article that many experimenters in the field of behavioral decision making and related areas in social and cognitive psychology tend to realize key variables of experimental design in a *fast* (e.g., using snapshot studies, and brief scripts), *inexpensive* (e.g., offering no financial incentives), and *convenient* (using deception) way. The drift toward these and other seemingly cost-effective experimental methods such as large classroom studies and take-home questionnaires (**Gigerenzer**) has occurred, we argue, due to a lack of both strong conventions and a theoretical framework which suggest how to implement experimental tests. While it is rational for experimenters as individuals to select methods and evolve conventions that minimize the costs (in time and money) of producing publishable data, we documented that this preference has a price tag that is too often overlooked: a greater likelihood of systematic data variability and error variance than alternative (and more expensive) methods would yield. Ultimately, the predominance of fast, inexpensive, and convenient methods of data collection is likely to contribute to a lack of replicability of experimental results. We identify the fast, inexpensive, and convenient route to data collection as one source of data variability. Experimental practices that contribute to this first source of data variability undermine control – "the hallmark of good experimental practice, whether it be undertaken by economists or psychologists" (**Harrison & Rutström**).

Another source of data variability in psychological experimentation is due to methodological diversity. Methodological diversity is high in the research areas we focused on because in these areas some researchers choose to implement more "expensive" realizations of the key variables, employing repeated trials and financial incentives and never using deception. The fact that many researchers use seemingly cost-effective methods, whereas others do not, is likely to

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

induce systematic data variability. The variability in empirical findings which we asserted and documented thus draws on two sources: the variability in results due to fast, inexpensive, and convenient experimental methods (what **Davis & Durham** call lack of "reliability" and what **Guala** calls lack of "clear-cut design") and due to the fact that a small but significant number of experimenters actually use other methods (what Guala calls "varied designs"). In the target article, we did not distinguish as clearly between these two sources of variability as, with thanks to our commentators' insights and our own hindsight, we now realize we should have. This unfortunate fact seems to have been the reason for some commentators' concern that we are out to stifle experimental diversity and creativity.

The do-it-both-ways rule (which accords key variables of experimental design the status of independent variables) does not post a risk to methodological diversity and experimental creativity for three reasons. First, the rule is tailored to the four key variables in question, and is not meant to interfere with other aspects of experimentation (i.e., our discussion has no bearing on the "participant-observer methodology or single case studies in clinical psychology" as **Davis & Durham** suggest). Second, in contrast to **Davis & Durham**'s, **Gil-White**'s, and **Maratsos**'s explicit reading and other commentators' (e.g., **Betsch & Haberstroh, Guala, Suleiman**) implicit suggestion, we do not endorse *empirically blind* rules such as economists' strict convention of always using financial incentives.[2] Rather, design and implementation decisions ought to be informed by the evidence rather than by beliefs, habits, or rituals. Third, the do-it-both-ways rule – applicable when evidence is unavailable or mixed – is a systematic reminder to implement more than one realization of a key design variable. It acknowledges that methodological variables represent auxiliary hypotheses (**Gillies & Rigdon, Smith**) and makes them an explicit part of theory testing. The do-it-both-ways rule broadens our experimental inquiry as it adds to researchers' methodological repertoire of fast, inexpensive, and convenient methods, alternative realizations of key variables (a consequence that **Maratsos** also foresees). Ultimately, the do-it-both-ways rule will counteract the de facto hegemony of the seemingly cost-effective methods that presently contribute to what we identified as the first source of data variability.

We admit that our suggestion to eschew deception, whenever possible, imposes constraints. We do not think, however, that such a convention undermines experimental ingenuity. The fact that deception – notwithstanding the APA's admonition to use it only as a last-resort – is still frequently used, indicates that there are no strong incentives to develop, evaluate, and employ alternatives. Making deception into a strategy of last resort is likely to spur the invention of new methods (as suggested by Baumrind 1971, and as exemplified by Bardsley 2000). We now turn to commentators' responses to our discussion of four key variables of experimental design.

R3. Enacting a script versus "ad-libbing"

There were apparently misunderstandings and questions about what we meant by a script. Economist **Roth**, for example, was "not sure that the use of 'scripts' is common in economics experiments." And psychologist **Baron**, for example, said, "I do not understand what is not a script . . . " What is a script, and why does it matter?

We defined a script as clear and comprehensive instructions which detail players (e.g., buyer, seller, market analyst, proposer, responder), their action choices, and the possible consequences of their choices (i.e., the payoffs). In addition, we described the particular kind of role-playing typically employed in economics experiments. Letting participants take on a particular role – having them enact a script – can be used to study not only strategic interactions but judgment, reasoning, and memory performance (e.g., Wason selection task, hindsight bias).

In our opinion, having participants enact explicit and comprehensive scripts has four potential advantages. First, scripts may constrain participants' interpretations of the situation by focusing their attention on those aspects that are intentionally communicated by the experimenter. The hindsight bias studies we described illustrate this point. Davies (1992) only told participants to recreate a previous state of knowledge, thus leaving participants to decide whether they should (1) attempt to retrieve their previous judgment as accurately as possible, (2) look as good (i.e., knowledgeable) as possible, or (3) spare their cognitive effort as their recollections would have no tangible consequences. In contrast, in Camerer et al.'s (1989) market experiment, the objective of avoiding the hindsight bias followed from being a successful trader; in other words, the role per se clarified the demand characteristic of the situation. Second, scripts can promote participants' active involvement in the experiment by making their choices ones that have tangible consequences. Third, scripts (especially if they are not abstract) restrict the sets of perspectives that participants bring to a particular judgment, choice, or reasoning task, and thus allow us to explore the mechanisms by which perspectives affect task performance (e.g., inference mechanisms as suggested by Cosmides & Tooby 1992 or (a)symmetric loss functions as suggested by **Weber**).

Finally, explicit and comprehensive scripts are the basis on which the sometimes subtle influence of instructions can be studied. There is, for instance, intriguing evidence that seemingly tiny procedural differences can make a large difference to behavior. Recently, Burnham et al. (2000) and Hoffman et al. (2000) showed, for instance, that changing the word "opponent" to "partner" or prompting players to think strategically before making an offer can have significant effects on how they behave in various contexts (for another striking example, see Harrison 1999, pp. 26–28). Therefore, explicit and comprehensive instructions enhance procedural regularity and ultimately, we claim, replicability (see also Binmore 1999). Scripts are thus one key to understanding the variability of experimental results – in economics and psychology. For an example, consider the test of the ultimatum game which one of our commentators, **Henrich**, conducted.

Whereas previous tests demonstrated that the "normative" solution of an uneven split was not a good description of empirical results, Henrich (2000) found that the Machiguenga people of the Peruvian Amazon make decisions which are much closer to the game-theoretic prediction. Henrich used instructions ("a set script"), but he had to explain the game at least three times. In addition, "often numerous examples were necessary to make the game fully understood" (p. 975). The experiment itself was introduced to an initial group of participants "under the auspices of 'play-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

ing a fun game for money'" (p. 975). Whereas Henrich (2000) suggests that "procedural differences seem unlikely to explain the substantial differences in observed behavior" (p. 975), Burnham et al.'s (2000) and Hoffman et al.'s (2000) results suggest the opposite. Certainly, the specific aspects of design mentioned above represent significant departures from the standard scripting of ultimatum games. For example, there is no telling what the impact was of the repeated explanations of the set-up or the numerous (possibly unscripted) examples. There is also a good chance that the framing of the experiment as a "fun game" being played for money hadan impact on the result. While there are still other possible explanations for the surprising results (e.g., the relative social distance among Machiguenga families), we argue that a clear and comprehensive script, to which the experimenters religiously adhered, would have increased one's confidence in the robustness of the reported results.

In the context of scripts, **Suleiman** points out that the "clarification of the demand characteristics is unavoidably entangled with the enhancement of those demand characteristics which coincide with the experimenters' focus." We agree and share his concern – as our comment on the reiterated explanations to participants of the ultimatum game in Henrich (2000) illustrates. We also share Suleiman's assessment of a problematic instruction from a well-known economic experiment published in 1985 – although the really interesting issue related to this two-stage implementation of the ultimatum game is, as one of the authors of the incriminated sentence has pointed out elsewhere, "why did [participants] not do what they were told at the first trial?" (Binmore 1999, F20). That said, we suggest that the benefits from clear and comprehensive instructions typically outweigh the costs of demand effects. Of course, even this is an empirical issue and can be approached as such. Relatedly, **Goodie** makes the excellent point that "one does not know if scripted interactions are representative of nonscripted ones." To which we say, amen, and would add that one also does not know whether nonscripted interactions are representative of that which an experimenter would like them to be representative. Again, even that is an empirical issue.

None of this implies (as, for example, **Davis & Durham** and **Huettel & Lockhead** intimate) that we try to exorcize incomplete information or uncertainty (including expectations regarding the behavior of other participants) from experimental settings. The provision of clear scripts and the systematic manipulation of (incomplete or uncertain) information are not mutually exclusive. Economists, for instance, routinely and systematically try to understand how "a choice by one person is affected by her or his expectations about the choices that might be made by other people" (Huettel & Lockhead). As a matter of fact, the recent literature on various bargaining games (e.g., Charness & Rabin 2000; Dufwenberg & Kirchsteiger 2000; Falk & Fischenbacher 1999) is all about this issue. Our point is that scripts can help to reduce unwanted, uncontrolled, and unnecessary uncertainty by channeling participants' interpretations of the experimental situation. Giving as little information as possible to participants about roles and perspectives as **Van Vugt** advises does not mean that participants will not bring perspectives to a particular judgment or choice task – only that no attempt has been made to control them or to understand how they affect task performance.

Regarding the provision of scripts (and the realization of other key variables), **Weber** emphasizes that existing theory can and should guide experimenter's decisions. We agree: The appropriateness of a key variable's implementation depends crucially on what aspects of behavior and cognition the theory under test is designed to capture. We also agree with Weber that, ideally, conflicting results should be resolved by theory rather than ever more experiments. When such a theory does not exist or is based on little empirical support, however, we suggest that the "ecumenical" do-it-both ways rule is a good idea. Being admittedly costly in the short and medium term, it promises to be cost-effective in the long term. By employing the rule, researchers are likely to accrue knowledge that allows us to eventually resolve seemingly neverending debates (about, for instance, whether or not incentives matter) and thus to allocate our resouces more efficiently in the future.

To conclude: Of the four key variables of experimental design, the effects of scripts (and their enactment) are most difficult to analyze due to scripts being rarely treated as independent variables. Thus, the evidence for our claim that providing a script affects results obtained is tenuous. While psychologists and economists are likely to share the view that supplying clear and comprehensive instructions is good experimental practice, there is nevertheless a difference between a script that details a role and scant instructions which make no reference to a role or a perspective (which, to answer **Baron**, counts as no script). Does the difference matter? We delineated four reasons why it might matter. Fortunately, whether it actually does can (and, we argue, should) be investigated experimentally.

R4. Repeated trials versus snapshot studies

No doubt, the study of *both* one-shot and repeated game and decision situations is useful (see Barron & Erev 2000). Indeed, economists have recently contributed models of (noisy) introspection for one-shot games (namely, Goeree & Holt 2000; in press b; Olcina & Urbano 1994; Stahl & Wilson 1995) that provide an explanation for departures from normative solutions.

Our argument is not that there is no place for experiments carried out only once, rather, that there has been little emphasis on repetition and feedback in psychological research (as, for instance, our analysis of the Bayesian reasoning literature illustrated). Why do we think that repetition and feedback are important?

R4.1. Why repetition?

We advocate repetition not only because the environment typically forces us into repeated decision and game situations (as suggested by **Betsch & Haberstroh**). Rather, our concern is that participants have a chance to become familiar with what is, under even the best of circumstances, an unusual situation. We gave these "practice effects" (**Baron**) as the first and foremost reason (for similar argument see Binmore 1999). In decision situations (i.e., "games against nature"), the particular kind of repeated trials which we discussed – stationary replication – means repeated decision-making or judgments in the same scenario. In game situations, stationary replication often takes the specific form of a "turnpike design" where one makes repeated decisions but

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

encounters the same player(s) only once. A second motivation for the use of repeated trials is specific to interactive situations. Repeated trials of this kind afford participants the opportunity to learn how their own choices interact with those of other players in a specific situation. We acknowledged that these two kinds of learning may be difficult to distinguish. Still, in the target article we pointed out

> the first kind of learning (adapting to the laboratory environment) relates to a methodological concern that participants may not initially understand the laboratory environment and task, whereas the second kind of learning (understanding how one's own choices interact with those of other participants) relates to the understanding of the possible strategic aspects of the decision situation.

R4.2. Stationary replication and other repeated trials

Gillies & Rigdon take us to task for our rationalization of the frequent use of repetition in economics. They argue that we get the history of game theory backwards. However, we did not aspire to tell the history of the eductive and the evolutive approach to equilibrium (selection) – one of us has done this elsewhere (Goodie et al. 1999) and has contributed to an analysis of the comparative advantages of the two approaches (Van Huyck et al. 1995; see also Blume & Ortmann 2000). Our point was simply to note that most economists are interested in equilibrium behavior and that experimentalists often justify their focus on play in late rounds in this manner (Camerer 1997).

Gillies & Rigdon also suggest that we do not understand that repeated games may generate additional equilibria. We did not focus on what Gillies & Rigdon call "repetition with replacement." Rather, we discussed stationary replication. In this context (i.e., when one employs a turnpike design), their objection (see also **Harrison & Rutström**) that repetition of the trust game was likely to lead participants to "update their estimates of the distribution of player types (trusters or not) in the environment" is well-taken. However, if experimenters are concerned about such an effect (as well they should be) they can always refrain from giving feedback until all rounds are played. It is possible that this would not get rid of the problem completely because there is evidence that simple repetition even without feedback has effects (e.g., Keren & Wagenaar 1987, or more recently Barron & Erev 2000) but we believe that simple repetition without feedback significantly attenuates the problem brought up by Gillies & Rigdon and Harrison & Rutström.

Henrich argues that another problem with repeated-game experiments is the almost complete emphasis on studying individual learning, as opposed to social learning. This, lest we misunderstand, strikes us as an untenable statement. Experimental economists routinely study, for instance, the emergence of conventions (e.g., Van Huyck et al. 1995; Young 1993). In fact, what is now called (evolutionary) game theory (e.g., Weibull 1995; see also Binmore 1994; and earlier Smith 1976[1759]) is all about social learning; so is much of game theory due to its interactive nature as **Ross** points out.

R5. Financial incentives versus no incentives

In response to our analysis of financial incentives and our "exhortation to incorporate decision makers' motivations more explicitly into judgment and choice models" (**Weber**), commentators focused on three major issues: the conditions under which incentives are (not) suitable; the difference between real and hypothetical payoffs; and the effect of financial incentive on anomalies (i.e., violations of normative standards) in individual decision making. We address each of these issues in more detail.

R5.1. When should incentives be used?

We suggested two *criteria* for the use of incentives: that research focus on people's maximal performance and that standards of optimal behavior be available. In addition, we proposed a simple decision tree to determine whether or not incentives should in fact be used when both criteria are met. First, is there evidence in past research regarding the effects of incentives? If "yes," does the available evidence indicate that financial (or possibly other) incentives affect behavior? If "no," we suggested applying a simple do-it-both-ways rule, thus according financial incentives the status of an independent variable.

In light of this approach, warning that relying exclusively on financial incentives would miss some important phenomena (**Roth**), or that investigating behavior such as child rearing using financial incentives would be inappropriate (**Davis & Durham**) are orthogonal to the policy we propose. Evidently, our policy does not adopt economists' current practices lock, stock, and barrel, nor does it define financial incentives to be the norm in decision experiments (as suggested by **Gil-White**). Moreover, the policy does not deny the exciting possibility that less effortful processes can outperform more effortful ones (**Betsch & Haberstroh**; see also Hertwig & Todd, in press) or that decision parameters differ across domains (**Rakow**). Of course, this approach also does not deny that incentives other than money may motivate participants (e.g., credit points; **Goodie**). In this context, it is heartening to see that even economists have started to explore the effects of financial incentives systematically, rather than taking them for granted. Schotter and Merlo (1999), for example, have translated the insight that exploring the strategy space may be (cognitively) expensive to participants in an experimental design which demonstrates that not paying participants while they learn can lead to significant improvements in outcomes (see also Berninghaus & Ehrhart 1998).

Betsch & Haberstroh discuss a set of four "principles" that must be met for financial incentives to be beneficial in terms of people's performance. The first "principle," availability of exact performance criteria, reflects common wisdom. Next, as the second and third "principles," they stress that "external"(experimenter's) and "internal" (participant's) criteria, together with participant's and experimenter's representation of the task, should be congruent. We understand that Betsch & Haberstroh do not presume such congruence to be necessary for financial incentives to be employed (as we can never a priori ascertain it and because payoff instructions are one key route to aligning experimenters' and participants' task representations) but rather that discongruence is a good candidate explanation when financial incentives have no effect or impair performance. As a fourth "principle" for incentives to be beneficial, Betsch & Haberstroh propose that a deliberate strategy must be the most appropriate way to solve the problem. They claim that people often use other than deliberate processes to

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

their advantage. This is a point well-taken. Previous research (in psychology and even more so in economics) has underestimated the role of simple, noneffortful, possibly automatic processes. That said, aspects of our memory, judgment, and decision-making performance are, as we documented, under strategic control, and thus the amount of cognitive effort invested can affect the performance. But even when less effortful processes are the objects of study, financial incentives can make the demonstration that those processes outperform effortful ones even more convincing. Moreover, financial incentives can help to explain where and how effortful processes can go wrong. Betsch & Haberstroh's conclusion that because financial incentives "might have counterproductive effects," they ought not to be used is a non sequitur. To rephrase Betsch & Haberstroh's conclusion, not taking the role of methodological key variables for granted but subjecting them to systematic variation paves the way to good experimentation (see **Fantino & Stolarz-Fantino, Guala, Harrison & Rutström**).

R5.2. Aren't hypothetical incentives sufficient?

Kühberger offers an intriguing argument: Since decision making involves anticipation of hypothetical events (e.g., future feelings, states of the world), hypothetical decisions are a valid proxy for people's real decisions. The implication of his argument is that hypothetical payoffs may suffice to study people's decisions. Kühberger, however, qualified his claim: The decision's importance can turn hypothetical decisions into invalid proxies for real ones. **Holt & Laury** precisely demonstrate such a case. In a choice between lotteries, they find comparable amounts of risk aversion for hypothetical and low real payoff conditions. A high real payoff condition, however, produced drastically different risk attitudes. Holt & Laury (personal communication) have since strengthened these results by conducting additional sessions with even higher real payoffs. Relatedly, **Fantino & Stolarz-Fantino** suggest that the importance of what is at stake may also explain why experiments with humans obtain less impulsive behavior than experiments with pigeons. The incentives experimenters offer to pigeons dwarf those offered to humans (of course, even here Harrison's [1994], dominance critique of payoffs might apply).

Although **Kühberger** acknowledges that what is at stake matters and **Holt & Laury** demonstrate that high payoffs can cause dramatic differences, they arrive at opposite policy recommendations. The former stresses the need for a theory of when and why real and hypothetical decisions coincide. In the absence of such a theory, he considers the do-it-both-ways rule a waste of money. In contrast, Holt & Laury argue that even while the discipline lacks an accepted theory of when financial incentives matter, they should nonetheless be used in economics experiments. If they have in mind a convention of always using financial incentives, we disagree. If, however, their argument is intended as a call to manipulate the provision of incentives systematically (as they in fact did), we agree. In contrast to Kühberger, we consider the do-it-both-ways rule (which in the present context may be better called the do-it-*n*-ways rule) an investment that promises high payoffs. Waiting for a theory of hypothetical and real decisions to emerge from an empirical vacuum seems overly optimistic. In fact, a comprehensive collection of reliable effects of financial incentives (as would quickly evolve if the do-it-both-ways rule was applied in psychology *and* economics) may act as a strong incentive to develop such a theory.

Finally, let us stress that the domain of high-stake decisions is not the only one where real and hypothetical incentives can yield divergent results. The results reported in Table 2 (of the target article) and in Camerer and Hogarth (1999) demonstrate that the payoffs need not be high stake incentives to affect people's judgment and decision-making.

R5.3. Do incentives eliminate anomalies?

From the 1970s, psychologists conducting research in the tradition of the heuristics-and-biases program have accumulated experimental evidence that suggests "behavioral assumptions employed by economists are simply wrong" (Grether 1978, p. 70). One prominent response of economists to this challenge has been to question the validity of the evidence. Experimentally observed anomalies ("fallacies," "biases," "cognitive illusions") could be, so the argument goes, peculiar to the methodological customs and rituals of psychologists (e.g., Grether 1980; 1992; Grether & Plott 1979). A number of commentators (e.g., **Fantino & Stolarz-Fantino, Gil-White, Holt & Laury, Kühberger, Zizzo**) take up this debate and discuss the robustness of anomalies. **Hilton**, for instance, asks whether anomalies can be eliminated by financial incentives and learning, and he concludes that the evidence strongly suggests that they cannot.

The picture is more differentiated and we agree with **Smith** that more "constructivity" is needed. It is time to go beyond blanket claims and categorical questions such as whether or not financial incentives eliminate anomalies. We also agree with **Fantino & Stolarz-Fantino**'s conclusion that the impact of key methodological variables on the results obtained "are not fixed and should not be taken for granted." **Holt & Laury**'s study is a good example of how differentiated the empirical pattern can be. In addition, it is important to acknowledge that conclusions regarding the effects of financial incentives (and similarly, repetition and feedback) are based on small and (sometimes opportunistic) samples of studies, and thus very likely are not the last word.

The little we know, however, suggests that financial incentives matter more in some areas than others (see Camerer & Hogarth 1999). Moreover, as we pointed out in Hertwig and Ortmann (in press), they matter more often than not in those areas that belong to the home turf of psychologists, namely, studies on judgment and decision making. Ironically, they may matter less in "game and market experiments" (Camerer & Hogarth 1999, but see Smith & Walker 1993a and Schotter & Merlo 1999), the home turf of economists. The need for a theory of the effects of financial incentives is apparent. We suggest that Camerer and Hogarth (1999) is an excellent point of departure. Most important, their capital-labor framework of cognitive effort highlights the interaction effects between key design variables such as repetition and financial incentives. Thus, it may have the potential to account for the conflicting observations regarding the effects of financial incentives.

Related to the existence of anomalies, **Gil-White** asks "if (1) people are given rewards for being good Bayesians, and (2) they receive feedback that is immediate and highly accurate, should we – upon the observation of results consistent with Bayesian reasoning – conclude that we have shown

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

that people are good Bayesians, or that experiments set up in this way can train them to be such?" The "context dependency of results" (**Levine**) is, indeed, an important question – but it is one that deserves equal attention in studies without rewards and feedback. To rephrase Gil-White: If people are given no rewards, and if they have only one or a few chances for an answer in an unfamiliar context, should we – upon the observation of results inconsistent with Bayesian reasoning – conclude that we have shown that people are bad Bayesians? In our view, **Hogarth** answers Gil-White's challenge: He reminds us that when faced with context-dependent results, researchers need to theoretically clarify the conditions to which results can be generalized.

Finally, we stress that the robustness of anomalies is debated not only in economics but also in psychology. In the wake of this debate, rich empirical and theoretical work in psychology has evolved that attempts to explain when and why people's inferences obey or disobey certain normative standards (see e.g., Erev et al. 1994; Gigerenzer 1996; Hilton 1995; Juslin et al. 2000; Krueger 1998). Thus, focusing merely on learning and financial incentives overlooks, for instance, what is likely to be the most powerful tool to reduce and sometimes even to eliminate "blunders of probabilistic reasoning" (Tversky & Kahneman 1987, p. 90), namely, to present statistical information in terms of (natural) frequencies rather than probabilities (e.g., Cosmides & Tooby 1996; Gigerenzer 1991a; Gigerenzer & Hoffrage 1995; Hoffrage et al. 2000). The seemingly robust conjunction fallacy (that **Zizzo** discusses), for instance, can be reduced and is sometimes completely eliminated when the information is presented in terms of frequencies (see Hertwig & Gigerenzer 1999, but also Mellers et al., 2001). Unfortunately, recent reviews of psychological literature for economists seem to be blissfully unaware of these empirical findings and theoretical discussions (e.g., Rabin 1998).

R6. Honesty versus deception

The use of deception in experiments entails costs. In light of the still frequent use of deception in some areas of psychology, it was surprising for us to see that none of the commentators explicitly question this assessment, and some commentators explicitly agree with it (e.g., **Baron, Goodie, McKenzie & Wixted, Zizzo**). We first clarify the definition of deception, then discuss McKenzie & Wixted's illustration of how suspicion contaminates experimental results, and finally explore when and why some commentators consider deception to be necessary.

R6.1. Reprise: What is deception, and what is not?

Baron proposes that not informing participants of the purpose of the study is "deception by omission" – but is it? In contrast to **Baron**, most researchers do not seem to regard the withholding of information as deception. Such agreement is, for instance, manifest in studies that review how often deception is used in psychological experiments. In Hertwig and Ortmann (2000), we examined the criteria for deception in those review studies. Intentional and explicit misrepresentation, that is, provision of false information, is unanimously considered to be deception. In contrast, not acquainting participants in advance with all aspects of the research being conducted, such as the hypotheses explored (e.g., the relationship between positive rewards and mood, to use Baron's example) is typically not considered deception. This view is also shared across disciplinary boundaries as the following statement by Hey (1998) illustrates: "There is a world of difference between not telling subjects things and telling them the wrong things. *The latter is deception, the former is not*" (p. 397).

Despite such a consensus, we appreciate **Baron**'s argument that withholding information makes full disclosure, when it is desirable, appear suspicious. In Ortmann and Hertwig (2000), we argued that one specific kind of "deception by omission" has the same potential for creating distrust as providing false information, namely, the violation of participants' default assumptions. For instance, a default assumption participants are likely to have is that a study starts only after an experimenter has clearly indicated its beginning. As a consequence, a participant might assume that her initial interactions with the experimenter (upon entering the laboratory) are not part of the experiment, and might feel misled if she finds out otherwise. We propose that violating default assumptions should be avoided.

R6.2. How participants' suspicions systematically contaminate data

McKenzie & Wixted provide two intriguing examples of how participants' distrust (likely to be fueled by the use of deception) systematically contaminated experimental results. Specifically, they show that the failure to recognize that participants may distrust experimenters' assertions about important task parameters (e.g., that a particular piece of information was randomly drawn) can lead participants' responses to be misclassified as irrational (e.g., as non-Bayesian). McKenzie & Wixted's analysis shows that participants' distrust has the potential of *systematically* distorting empirical observations and thus leading experimenters to draw erroneous conclusions – for instance, regarding people's ability to reason in accordance with normative principles (e.g., Bayes's rule). The threat of systematic contamination due to distrust has also been documented in other domains. In an extensive search for studies exploring the contaminating effects of deception, Hertwig and Ortmann (2000) and Ortmann and Hertwig (2000) found that, across a variety of research domains, personal experience with deception can and does distort observed behavior (e.g., judgments, attitudes, and measures of incidental learning and verbal conditioning).

As an option to deal with participants' skepticism about task parameters, **McKenzie & Wixted** propose incorporating a "trust" (or distrust) parameter into models of participants' behavior. While elegant, this approach introduces a free parameter into the models (thus increasing the danger of data-fitting, in particular when unlike in McKenzie and Wixted's models, more than one free parameter is involved). Moreover, we fear that this modeling approach will often not be applicable, as it demands a good understanding of where and how distrust interferes with the processes under consideration.

R6.3. Is deception indispensable, and is it treated as a last-resort strategy?

The most common argument for deception is that it is indispensable for the study of those facets of behavior for

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

which participants have reasons to conceal their truthful opinions, attitudes, or preferences. Therefore, experimenters must lie in order to avoid being lied to. Several commentators reiterated this argument (e.g., **Davis & Durham; Van Vugt**) or variants of it (**Baron**). In the absence of strong incentives to develop alternatives to deception, this rationale can only be evaluated in the abstract. Clearly, at this point there is no principled argument that could prove this rationale wrong. Consequently, we stated in the target article that we do not exclude the possibility that there are important research questions for which deception is indispensable. Irrespective of this issue, however, we argue that the prevalence of deception could substantially be reduced if it were used as a strategy of last resort. It is here where we disagree with **Weiss** who whole-heartedly defends the current practices.

Weiss complains that our focus on the methodological costs of deception is an ethical argument in disguise. In suspecting an ethical argument, Weiss refers to our public good analysis of deception. Clearly, our argument is orthogonal to the deontological arguments put forth by Baumrind (1964; 1979; 1985). We agree with Weiss that this analysis *invites* a value judgment, namely, that the experimenters who produce the public good while others do not are being exploited. But, such a judgment surely does not make the analysis of deception in terms of a public good problem less valuable; nor does it absolve the defenders of deception from discussing its contaminating potential. Weiss does not devote a single word to this potential.

Ironically, **Weiss** himself points out that deception comes at a considerable cost, namely, that of an arms race in which experimenters have to design even more sophisticated ways of camouflaging the true purpose of the experiment since participants may become increasingly distrustful and sophisticated in figuring out where deception occurs.[3] He compares behavioral researchers to medical researchers who have to continuously invent new antibiotics (or novel techniques of deception) to keep abreast of the microbes (or participants' suspicions). Why? Because, and for us this is the crucial lesson of Weiss's analogy, they are often not used as a treatment of last resort but instead as first-resort treatment in cases of common infections (or simply to promote growth in animals; see *The New York Times*, January 17, 2001, Section: Living). The unfortunate consequence is that bacterial resistance to antibiotics is thus much more likely to evolve. Take the U.S. and the Netherlands as examples. In the U.S., it is routine to use antibiotics for treating middle ear infection, one of the most common diagnoses in children, whereas in the Netherlands, the standard practice is to use antibiotics only if the infection fails to improve after a time of 'watchful waiting.' Not suprisingly, bacterial resistance in the Netherlands is about 1%, compared with the U.S. average of around 25% (see http://healthlink.mcw.edu/article/965945751.html).

Just as antibiotics (in the U.S.), deception is not exclusively used as a last-resort strategy. In contradiction to **Davis & Durham's** belief, even a cursory glance at contemporary deception studies reveals that deception is used even when it is not indispensable (recall that every third study in *JPSP* and every second study in *JESP* uses deception). Examples include studies in which participants are instructed that incentives are performance contingent when in reality they are not or the claim that some aspect of an experiment (e.g., the allocation of a specific role, a piece of information, etc.) was randomly chosen when in reality it was not (for more details see Hertwig & Ortmann, in press). Deception is a method that saves resources (as **Baron** points out) but it is only inexpensive if there will be no (or only minimal) costs for future experiments. But are there really no costs? We doubt that the participants believe promises of performance-dependent rewards at face value in future experiments if they just found out (through debriefing) that the experimenter misled them on the contingency of those rewards. Once bitten, twice shy.

The evidence regarding the consequence of firsthand experience with deception (see Hertwig & Ortmann 2000; Ortmann & Hertwig 2000) counsels us to treat deception as a last-resort strategy, thus limiting the number of participants with firsthand experience. In fact, this is the policy as currently stipulated by the APA guidelines. Considerations as formulated by **Baron** (resource saving) and **Van Vugt** (deception is justified when beneficial in the development of nontrivial theories of human behavior) are not endorsed by the APA guidelines. Finally, experiments that a number of commentators (e.g., **Davis & Durham, Hilton, Roth**) consider to be prime examples of cases in which deception was indispensable or yielded valuable, and eminently teachable, insights into human behavior would likely not pass contemporary ethical committees (e.g., the Milgram experiment). Therefore, the utility of those studies does not absolve us to/from deciding anew about the use of deception in present experiments.

How can deception be implemented as a last-resort strategy and how can the existing distrust among participants be overcome? In Hertwig and Ortmann (2000), we propose an incentive-compatible mechanism that has the potential to reduce the frequency of deception (and to promote methodological innovation). To overcome existing distrust, each individual laboratory can attempt to (re-)establish trust by taking measures such as introducing a *monitor* into experimentation (i.e., participants elect one of themselves to be a paid monitor who inspects all equipment and observes all procedures during the experiment; see Grether 1980; 1992; for a similar proposal see **Baron**). Such concrete gestures to (re-)gain participants' trust may also help to shorten the time any policy change will require for psychologists to overcome their reputation (a problem that **Roth** points out).

As still another remedy, **Henrich** proposes conducting deception studies outside the laboratory. Albeit an original proposal, we doubt its long term utility. If psychologists restricted the use of deception to studies done outside the university, this practice would quickly become public knowledge. Such knowledge, and the expectations it is likely to evoke, may compromise not only the work of researchers who conduct field studies but also that of professional psychologists in general.

R7. Additional issues: Subject pools, institutional arrangements, and the art of fast data collection

In reflecting on how we go about our business, several commentators highlighted practices that, in their and our view, deserve closer scrutiny. **Henrich**, for instance, criticizes the reliance of both psychology and economics on university students as participants – a "very weird, and very small, slice

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

of humanity." He argues that as a result of this practice, researchers from both fields overgeneralize their results.

Psychology's reliance on a highly selected subject pool may be even more pronounced than Henrich assumes. According to Sieber and Saks (1989), "undergraduate students have been a major source of research data in many areas of psychology" (p. 1053). The participation of undergraduates in the subject pool is typically institutionalized through the requirement that students in introductory psychology need to participate in (some) research projects as part of their course requirements. This availability of "free subjects" may be a key to understanding psychologists' experimental practices. Vernon Smith (personal communication) once asked a colleague from psychology "why psychologists, who you would expect to be concerned about motivation, did not pay salient rewards to subjects. [The psychologist] said it was simple. Every psychology department requires majors to participate in a minimum number of experiments as a condition for a degree, and that it was unthinkable that you would start using rewards for this huge population of free subjects."

While the use of deception appears to have immediate detrimental effects – namely, suspicion and distrust among those who experienced it – the institutional response to those effects may also come with a price tag: There is evidence indicating that since the 1960s the proportion of participants in psychology experiments from introductory courses has been on the rise (see Hertwig & Ortmann, in press). This change to the current widespread recruitment from introductory courses (and thus reliance on an even more unrepresentative subject pool) can be read as an institutional response to the risks of a subject pool contaminated by distrust (Ortmann & Hertwig 2000). Recruiting participants who are less likely to have firsthand experience with deception – students in introductory classes – minimizes the problem of participants' suspicions.[4] Albeit speculative, this explanation is consistent with the advice psychologists were given not to use the same (deceived) students twice: Based on experimental evidence, Silverman et al. (1970) concluded, more than 30 years ago, "that the practice of using the same subjects repeatedly be curtailed, and whenever administratively possible, subjects who have been deceived and debriefed be excluded from further participation" (p. 211). Increased reliance on introductory students has – to the best of our knowledge – not been observed in economics.

In our view, **Hogarth** suggests one important step toward a systematic remedy of a potentially unrepresentative subject pool. He calls for theoretical clarity about the kinds of people and tasks to which the results might be generalized. In addition, he advocates the combination of laboratory and field studies, thus rendering it possible for experimenters to explore the generalizability of their laboratory results. In Hogarth's view, economists more than psychologists explicate the characteristics of people and tasks to which experimental results are meant to be generalized. By attending to the task, (and its correlates in the real world) Hogarth reminds psychologists of a concern that is reminiscent of Egon Brunswik's (1955; as does **Weber** by emphasizing "compatibility" between the experiment and its targeted real-world application). Brunswik criticized his colleagues for practicing "double standards" by generalizing their results to both a population of situations and a population of people, although only being concerned with the sampling of the latter.

Gigerenzer describes a tendency in psychological experimentation to conduct large-scale data collection outside the laboratory context, in particular, the practices of "take-home questionnaires" and "large classroom experiments." These methods are another instance of fast and seemingly inexpensive methods. Gigerenzer suggests, that their use at American universities may help to explain repeatedly observed performance differences between American and German studies examing probabilistic and logical reasoning. If so, these methods may also help explain why much past research in behavioral decision-making has arrived at rather pessimistic conclusions regarding people's reasoning abilities. Clearly, this hypothesis can be tested empirically.

R8. Why do the methodological practices differ?

A number of commentators argue that the answer simply is that practices (must) differ because the two disciplines have different subject matters and research questions (e.g., **Davis & Durham, Gil-White, Lecoutre & Lecoutre, Suleiman**). But is it "really the question being asked that will always determine the methodology" (**Davis & Durham**; see also for a similar form-ought-to-follow-function argument, **Baron, Huettel & Lockhead, Lecoutre & Lecoutre, Roth**)? **Gigerenzer** reveals the ahistorical naivete of this claim. Even within psychology, this "naturalistic" argument fails to explain surprising structural similarities in different theoretical traditions. What are more substantial explanations of the observed differences? **Blaich & Barreto** suggest that the different practices may be driven by economists' and psychologists' different use of statistics. **Gintis** reiterates the conjecture in our target article of the unifying role of game theory, arguing that the emergence of game theory suffices to explain the differences in experimental practices. Finally, **Huettel & Lockhead** argue that psychologists' and economists' experiments serve different functions.

Before proceeding to address these proposals, let us distinguish two aspects of methodological standards, namely their "nature" and their "content." The former refers to how binding standards are; the latter to their actual substance (e.g., use of financial incentives). We suggest that any explanation of the different nature of standards in economics and psychology needs to acknowledge the broader context from which experimentation emerged in either field. **Hogarth** and **Roth** reiterate our suggestion that experimental economists had to fight hard to be accepted in a profession that for many years doubted the utility of laboratory experiments for making inferences about the real world (as can be glimpsed from the introductory remarks of two path-breaking papers, Smith 1976; 1982). In contrast, experimental psychologists never similarly had to battle for respect within their own discipline. While the specific circumstances of their emergence may explain why methodological standards are less binding in psychology than in economics, history does not explain the content of the standards. How did the content evolve?

R8.1. Use of statistics

In **Blaich & Barreto**'s view, research practices differ because psychologists and economists make different use of inferential statistics: "The fact that experimental psycholo-

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

gists tend to assign much more importance to rejecting the null hypothesis but less importance on making precise parameter estimates than experimental economists plays an important rule, in our view, in creating the differences in the two fields." Although we are not convinced that these differences have caused different methodological practices (e.g., psychologists' use of deception and economists' proscription of deception), it may very well be that psychology's practice of null-hypothesis testing perpetuates differences. It does so by impeding the elaboration of precise theories (an argument that has repeatedly been made within psychology, Hertwig & Todd 2000; Krueger 1998; Schmidt & Hunter 1997). Imprecise theories, in turn, tend to leave decisions on how to realize experimental tests to the discretion of the researchers, and thus to the dominant methodological preferences in a field. In contrast, precise theories are more likely to imply appropriate test conditions, for instance, by explicitly defining the behavior it targets (e.g., first impression, learning, equilibrium behavior).

R8.2. The role of game theory

Gintis claims that there is a simple answer why methodological practices differs. It is because economists use game theory to design and interpret experiments. Although we hinted at the unifying role of game theory, its presence cannot explain why methodological conventions have such a regulatory nature in economics. We believe that the most plausible candidate explanation for their nature is the strategic role that the canonization of mandatory rules played in the process of gaining acceptance within the discipline. With regards to the content of the conventions, **Gillies & Rigdon** argue – contrary to **Gintis**'s thesis – that three of the four key variables (namely, deception, financial incentives, and scripting) are "general" variables "whose appropriateness is independent of the theory being tested." Whether or not this assessment is correct, we anticipate that any comprehensive explanation of why methodological practices in the two fields differ will involve among others, an understanding of the role of early key players (**Holt & Laury** and **Smith** point out that a psychologist, Sidney Siegel, has been largely responsible for establishing the procedural standards used in economics experiments[5]), the role (or relative lack thereof) of unifying theories (e.g., game theory, behaviorism), institutional arrangements (e.g., the availability of subject pools) as well as the fact that experimental economics for a significant number of years was done only in about half a dozen laboratories.

While discussing game theory, let us note that **Van Vugt**'s assertion that "economists [and hence experimental economists] use the theory of self-interest as [a] unique explanatory framework for understanding human behavior," is wrong. It demonstrates lack of awareness of the theoretical developments that have dramatically reframed economics – primarily by way of mostly game-theoretic reformulations. We doubt that there are economists out there who do not believe that "given the right conditions, people can be rational or irrational, selfish or altruistic, aggressive or helpful" (Van Vugt). We are certain that if indeed such an "econ" (Leijonhufvid) exists, he or she is not an experimental economist. Van Vugt's wage escalation argument, furthermore, misapplies basic tenets of marginal utility theory. Money is typically chosen because of its, for all practical purposes, nonsatiation property. By Van Vugt's logic, real wages would go up and up and up. . . . Last but not least, and also regarding issues of homo economicus, **Zizzo** takes us to task for our comments on Frey's work on intrinsic motivation. We urge the reader to re-read footnote 8 of the target article and read the references therein.

R8.3. Do psychologists generate questions, whereas economists test models?

Huettel & Lockhead make a distinction between "restricted experimental designs, which allow reproducibility and hypothesis testing, and exploratory designs, which may provide new insights into phenomena." Based on this distinction, they suggest that economics studies were designed to answer well-defined hypotheses, whereas the psychology studies in question have more of an exploratory character.

Huettel & Lockhead's characterization of the psychological studies in question is not well-informed. Research in experimental economics and psychological research on judgment and decision making are particularly well suited for a comparison of methods across disciplines because studies in both fields often address similar and sometimes even identical questions. As examples, consider questions such as whether or not people update probabilities in a Bayesian way, make choices in a transitive way, are subject to the hindsight bias ("curse of knowledge"), or allocate resources in a way that satisfies rational economic theory (or motives such as fairness). Sometimes economists and psychologists explore exactly the same hypothesis: for instance, whether or not people apply the representativeness heuristic (Kahneman & Tversky 1973) to update probabilities (e.g., Grether 1980; 1992; Harrison 1994). Arguing, as Huettel & Lockhead do, that the economists' and "the particular psychology studies that were selected for consideration" differ because the latter are in the business of "generating questions," whereas the former test well-defined hypotheses reveals, depending on the perspective, a rather self-deprecating or condescending attitude. We do, however, agree with **Hogarth**'s assessment that many psychological studies test theoretical notions rather than formalized theories or process models – exactly this fact has been at the heart of a controversial debate among psychologists (Gigerenzer 1996; Kahneman & Tversky 1996).

Despite our disagreement with how **Huettel & Lockhead** characterize the goal of psychological studies, we appreciate the more general question they raise, namely, whether or not our policy recommendations should (equally) apply to hypothesis-testing and hypothesis-generating experiments. While we agree with Huettel and Lockhead that in the context of discovery "everything goes," we point out that placing participants in a not well-defined situation is only one tool and probably not a particularly productive one for generating questions (for other tools used in the context of discovery see, for instance, Gigerenzer 1991b).

In the context of theory testing, **Erev** highlights one of the crucial benefits of standardized test conditions, namely the emergence of data sets that, because of being collected under comparable conditions, can be used *in toto* to test a hypothesis, a model, or a theory. Such a body of data will allow researchers to use a strict rule, the "generality first" rule, in the process of theory selection. This rule requires that a new model replaces an old model only if it explains previous data plus new data that the old model cannot accommodate. While we suggest that this rule should not be

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

used in isolation (but be complemented by other theory-selection criteria such as internal consistency and simplicity), we agree with Erev that the evolution of large standardized data sets is one promising route to cumulative progress in modeling. We also agree with Erev that the do-it-both-ways rule will first quickly help to identify how key variables of experimental design affect the results obtained, and then, once such knowledge is available, will promote the evolution of data sets collected under comparable conditions.

As an aside, the generality-first rule also implies a third way of testing theories – beyond null-hypothesis testing and parameter estimation (**Blaich & Barreto**). According to this rule, a theory is tested against the aggregate set of data, and its status (rejected/accepted) is a function of its explanatory power (regarding this set) and the performance of its competitors. Because they are intended to be useful approximations (**Roth**), theories can overcome rejection based on individual experiments if they still succeed in accounting for a wide range of observations.

R9. Experimental economics and behaviorism: Guilty by association?

> I have not the slightest doubt that if Sid Siegel had lived, say another 25–30 years, the development of experimental economics would have been much advanced in time. He was just getting started, was a fountain of ideas, a powerhouse of energy, and had unsurpassed technique and mastery of experimental science. Twenty-five years later I asked Amos Tversky, "Whatever happened to the tradition of Sidney Siegel in psychology?" His answer: "YOU'RE IT!"

In relating this episode to us, Vernon Smith made it clear that Tversky's response was meant to be a put-down. Tversky saw Siegel as one of the last of the Skinnerian behaviorists. Likewise, a number of commentators remark on what they see as similarities between the experimental practices of economists and behaviorists. **Fantino & Stolarz-Fantino** see this similarity in a positive light and illustrate how classic effects observed in the heuristics-and-biases program (e.g., base-rate neglect, conjunction fallacy) can be studied using methods from the learning tradition. For **Hilton** and **Kurzban** (and in a somewhat related way also **Maratsos**) in contrast, this similarity is a reason for concern. Admittedly simplified, Hilton and Kurzban's arguments are the following: First, the experimental methods in economics resemble those employed by behaviorists. Second, the methodological similarity indicates a theoretical affinity, with economists being "methodological behaviorists" who focus on observables at the expense of cognitive processes (Hilton; see also **Markman, Rakow**), or focus, like behaviorists do, on domain-general nonsignificant learning mechanisms. Third, either focus is a theoretical cul de sac, and "psychologists did the right thing to abandon behaviorism" (Hilton), whereas adopting economists' methodology in psychology would be tantamount to "behaviorist-like experiments" and akin to a return to the "dark days of behaviorism" (Kurzban).

We disagree with **Hilton**'s and **Kurzban**'s view. They seem to suggest that taking into account realizations of key variables of experimental design that economists and behaviorists value, goes along with adopting their imputed theoretical biases (i.e., focus on output or nonsignificant learning mechanisms). As **Gigerenzer** explains, however, there is not such automaticity. Even within psychology, there are research programs that are utterly different in their theoretical nature despite commonalities in experimental practices. Thus, even if it were true that economists focus on observable outcomes (at the expense of processes) as Hilton and **Rakow** suggest, nothing in the emphasis on learning and motivation excludes the study of processes (as, for instance, Wallsten's 1972; 1976, studies on probabilistic information processing illustrate). On the contrary, the provision of financial incentives is one important tool for decreasing variability, thus increasing the reliability of processes (and process measures); and the use of repeated trials is a powerful tool for studying the evolution of processes. There is hardly an automatic contingency between the use of financial incentives, scripts, and repetition, and the sudden disappearance of cognitive processes in a black box.

But is the conventional wisdom that **Hilton** and **Lecoutre & Lecoutre** express even accurate – that psychologists are process-oriented, whereas economists focus on observable outcomes? There are important counterexamples. Take, for instance, the most influential research program in psychological research on behavioral decision making, the heuristics-and-biases program. It seems fair to conclude that this program has an explicit focus on observable outcomes (**Markman** seems to agree). Compared to the search for new "biases," "fallacies," and "cognitive illusions," the modeling of the psychological processes has received little attention (see the debate between Kahneman & Tversky [1996] and Gigerenzer [1996]). Counterexamples to economists focus on observable outcomes are, for instance, the research programs by Camerer et al. (1993), Costa-Gomes et al. (2001), or McCabe et al. (2000). Whereas these researchers are still interested in outcomes, they focus on the reasoning processes underlying choices leading to outcomes, and even their neurological correlates (**Smith**).

Finally, what about **Kurzban**'s argument that economists (and behaviorists alike) study domain-general mechanisms of nonsignificant relevance? Although we are not sure what mechanisms Kurzban has in mind, it is worth remembering that theorizing about domain-specificity (as evolutionary psychologists such as Kurzban do) apparently can profit from domain-general theoretical frameworks such as game theory. Take Cosmides and Tooby's (1992) social contract theory, one major theory in recent evolutionary psychology, as an example. In their view, the barrier to the evolution of social exchange is a problem that is structurally identical to the one-move Prisoner's Dilemma, and indeed Cosmides and Tooby (1992) used this game to refine their theory.

R10. Experimental practices in psychology: A challenge for economists?

Several commentators point out that the methodological dialogue between economists and psychologists must not be a one-way street (**Harrison & Rutström, Levine, Roth**). We whole-heartedly agree. Indeed, such a debate can work only if both sides are open to the input from the other camp. Admittedly, we focused in our treatment on those key variables where we believe psychologists can profit from comparing their experimental practices with those of experimental economists, the new kid on the block.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

We also pointed out, however, that the conventions and practices of experimental economists (that **Weber** facetiously describes as "method fascism") do not constitute the gold standard of experimentation, and that "a paper entitled 'Experimental practices in psychology: A challenge for economists?' may well be worth writing." To our minds, there is no doubt that **Harrison & Rutström** are right on the money when they argue: "Unfortunately, experimental economists have sometimes followed conventional practices with little thought about the consequences." We share their skepticism of "the popular use of 'lab dollars'" (as, incidentally, do Davis & Holt 1993, p. 29). More generally, we also stress that the do-it-both-ways rule is a significant departure from empirically blind conventions that experimental economists currently take for granted.

R10.1. Acontextualization, field referents, and framing effects

Harrison & Rutström also discuss the issue of field referents that participants may bring into the laboratory. Economists typically try to overcome the problem of such imported priors by acontextualization – stripping the experimental scenario and instructions of any reference to the real-world problem that may have motivated the scenario. For example, in principal-agent games most experimental economists label the employee the "seller" and the employer the "buyer" of unspecified goods or services. Sometimes they even omit these labels and call the employee (employer), say, "participant A" ("participant B"). Although acontextualization has the advantage of counteracting the problems of uncontrolled priors that participants bring into the laboratory (an issue that **Fantino & Stolarz-Fantino, Goodie,** and **Betsch & Haberstroh** also highlight), it has two clear drawbacks. First, the abstract context invites sense-making exercises on the part of the participants who might try to make the connection between the laboratory set-up and possible real-world correlates. Second, the abstract context may prevent participants from invoking the kind of inference routines that they use to navigate similarly structured real-world environments. We use the word "routines" here intentionally because, although we disagree with their claim about the scope, we agree with Betsch & Haberstroh's emphasis of the importance of less effortful processes.

Relatedly, **Hogarth** argues that "theory (in economics) specifies that different structural representations of the environment (e.g., framing of decision problems) should make no difference. . . . Context – however vaguely defined – is important to psychologists, but not to economists." Although that statement is not true in its generality (e.g., Andreoni 1995; Offerman et al. in press; Ortmann et al. 2000; or the previously mentioned work by **Smith** and his collaborators), there can be no doubt that psychologists are overwhelmingly more sensitive to how problem and information representation affects people's reasoning.

R10.2. Heuristics and how to select them

Harrison & Rutström highlight the selection of heuristics as a theoretical theme that unites the field of experimental economics and psychology. We agree whole-heartedly. More generally, we believe that in a world in which knowledge and mental resources are limited, and in which time is pressing, the study of real-world judgments and decisions require alternatives to traditional models of unbounded rationality and optimization. In a recent *BBS* précis, Todd and Gigerenzer (2000) described the framework of fast and frugal heuristics and placed the study of those heuristics within the context of bounded rationality (Simon 1990). Looking toward future work, Todd et al. (2000) delineate three major theoretical questions: Where do heuristics come from? How are they selected and how are they adapted to the decision and environment structure in which they evolve? Seeking to answer these and related questions can foster a further theoretical convergence. Although we are not sure that the search for boundedly rational heuristics is what **Levine** envisions when he talks about a "psychologically based economic theory," we agree with him that emotions – whether they function as stopping rules for search (Simon 1956) or in some other way – will be a crucial topic in any program of bounded rationality.

Theoretical and methodological issues are often linked. The study of heuristics is a case in point. Obtaining empirical evidence for the use of particular heuristics demands careful methodology because of challenges such as the flat maximum phenomenon and individual differences in their use (a source of variance that **Suleiman** stresses). The study of heuristics will require psychologists and economists to make methodological decisions closely related to those that we have discussed here – decisions about the structure of the decision environment (e.g., abstract vs. content-rich), the incentive landscape (e.g., favoring accuracy, speed, or other performance criteria), or the structure and kind of feedback (to study the evolution and learning of heuristics). We agree with **Markman** that psychology has much to offer in terms of techniques for studying online processing and heuristics. In fact, economists have already started using techniques such as MouseLab (e.g., Camerer et al. 1993; Costa-Gomez et al. 2001).

R11. Conclusion

> Methodological discussion, like spinach and calisthenics, is good for us . . . . (Paul Samuelson, p. 231)

Commenting on presentations by Ernest Nagel, Sherman Krupp, and Andreas Papandreou on methodological problems, Samuelson (1963) noted that while undoubtedly methodological discussion is good for us, it is not often practiced and thus may be, ultimately, inconsequential. We hope that Samuelson is wrong with his assessment and that **Harrison & Rutström** are right with their generous claim about the effect of our target article. There is hope. After all, more people today appreciate spinach and calisthenics (although they typically have fancier names for the latter).

We are convinced that a common language of scientific discovery and theory-testing, in addition to experimental practices grounded in empirical evidence, promise high payoffs. Ultimately, of course, these claims are an empirical question. We can say for ourselves – one being a psychologist, the other being an economist – that we found the conversation across disciplinary borders a rewarding (albeit not always easy) exercise. We urge others to follow suit.

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

NOTES

1. We use the term do-it-both-ways rule as a short-hand expression. Obviously, there are situations where more than two realizations of a variable will be explored.

2. We note that there is one important exception to that statement: The work Cummings and Harrison and their collaborators have done on hypothetical bias in contingent valuation studies (see Harrison 1999 for an excellent survey and discussion).

3. Parenthetically, we note that we believe this concern to be significantly more relevant than **Van Vugt**'s concern about participants being less likely to turn up again after having experienced deception once.

4. Students in introductory classes are also less likely to have already concluded a large number of psychology courses. Rubin and Moore (1971) observed that the number of psychology courses which students had completed—not the number of deception experiments in which participants recall having taken part—correlated with participants' level of suspicion.

5. Revisiting the experimental practices of early key players in psychology also reveals that today's principles of experimentation are not necessarily the ones endored by those who founded the field of behavioral decision making. Take, for instance, Ward Edwards (personal communication), who says that the following key principle (among others) guided his PhD thesis: "Since the vagaries of circumstances, especially of human circumstances, are unlimited in complexity, the inevitably-inadequate tools that one tries to use in understanding and dealing with them must be as bullet-proof as possible. That calls for realism in experimental design (real payoffs, for example; more precisely, realism of payoffs was sure to be, and in fact was, an interesting and important experimental manipulandum)."

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Abraham, F. D., Abraham, R. H. & Shaw, C. D. (1990) *A visual introduction to dynamical systems theory for psychology.* Aerial Press. [DSL]

Adair, J. G., Dushenko, T. W. & Lindsay, R. C. L. (1985) Ethical regulations and their impact on research practice. *American Psychologist* 40:59–72. [aRH]

Aitkenhead, M. & Dordoy, J. (1985) What the subjects have to say. *British Journal of Social Psychology* 24:293–305. [aRH]

Allen, D. F. (1983) Follow-up analysis of use of forewarning and deception in psychological experiments. *Psychological Reports* 52:899–906. [aRH]

Allison, S. T. & Messick, D. M. (1990) Social decision heuristics in the use of shared resources. *Journal of Behavioral Decision Making* 3:195–204. [aRH]

American Psychological Association (1992) Ethical principles of psychologists and code of conduct. *American Psychologist* 47:1597–611. [aRH]

Andreoni, J. (1995) Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110:1–21. [rRH]

Arkes, H. R. & Ayton, P. (1999) The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin* 125:591–600. [EF, ASGo]

Balzer, W. K., Doherty, M. E. & O'Connor, R. (1989) Effects of cognitive feedback on performance. *Psychological Bulletin* 106:410–33. [aRH]

Bandura, A. (1977) *Social learning theory.* Prentice Hall. [JH]

Bardsley, N. (2000) Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics.* (in press). [rRH, DJZ]

Bargh, J. A. (1996) Principles of automaticity. In: *Social psychology: Handbook of basic principles,* ed. E. T. Higgins & A. Kruglanski. Guilford Press. [TB]

Bar-Hillel, M. & Fischhoff, B. (1981) When do base rates affect predictions? *Journal of Personality and Social Psychology* 41:671–80. [aRH]

Baron, R., Vandello, J. & Brunsman, B. (1996) The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology* 71(5):915–27. [JH]

Barro, R. J. (1990) *Macro-economic policy.* Harvard University Press. [aRH]

Barron, G. & Erev, I. (2000) On the relationship between decisions in one-shot and repeated tasks: Experimental results and the possibility of general models. Manuscript, Faculty of Industrial Engineering and Management, Technion. [rRH]

Baumeister, R. E. & Sommer, K. L. (1997) Consciousness, free choice, and automaticity. In: *The automaticity of everyday life,* ed. R. S. Wyer, Jr. Erlbaum. [TB]

Baumrind, D. (1964) Some thoughts on ethics of research. After reading Milgram's "Behavioral study of obedience." *American Psychologist* 19:421–23. [arRH]

(1971) Principles of ethical conduct in the treatment of subjects: Reaction to the draft report of the Committee on Ethical Standards in Psychological Research. *American Psychologist* 26:887–96. [arRH]

(1979) IRBs and social science research: The costs of deception. *IRB: A Review of Human Subjects Research* 1:1–14. [arRH]

(1985) Research using intentional deception: Ethical issues revisited. *American Psychologist* 40:165–74. [arRH]

Beach, L. R. & Phillips, L. D. (1967) Subjective probabilities inferred from estimates and bets. *Journal of Experimental Psychology* 75:354–59. [aRH]

Beattie, J. & Loomes, G. (1997) The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty* 14:155–68. [aRH]

Beeler, J. D. & Hunton, J. E. (1997) The influence of compensation method and disclosure level on information search strategy and escalation of commitment. *Journal of Behavioral Decision Making* 10:77–91. [aRH]

Berg, J. E., Dickhaut, J. W. & McCabe, K. A. (1995) Trust, reciprocity, and social history. *Games and Economic Behavior* 10:122–42. [ASGi, aRH]

Berg, J. E., Dickhaut, J. W. & O'Brien, J. R. (1985) Preference reversal and arbitrage. *Research in Experimental Economics* 3:31–72. [aRH]

Bergin, A. E. (1991) Values and religious issues in psychotherapy and mental health. *American Psychologist* 46:394–403. [HPD]

Berninghaus, S. K. & Ehrhart, K. M. (1998) Time horizon and equilibrium selection in tacit coordination games: Experimental results. *Journal of Economic Behavior and Organization* 37:231–48. [rRH]

Betsch, T. & Fiedler, K. (1999) Understanding conjunction effects in probability judgment: The role of implicit mental models. *European Journal of Social Psychology* 29:75–93. [TB]

Betsch, T., Haberstroh, S., Glöckner, A., Haar, T. & Fiedler, K. (in press) The effects of routine strength on adaptation and information search in recurrent decision making. *Organizational Behavior and Human Decision Processes.* [TB]

Betsch, T., Haberstroh, S. & Höhle, C. (2000) Explaining and predicting routinized decision making: A review of theories. (submitted). [TB]

Betsch, T., Plessner, H., Schwieren, C. & Gütig, R. (2001) I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin* 27:242–53. [TB]

Binmore, K. (1994) *Playing fair.* MIT Press. [GWH, arRH]

(1999) Why experiment in economics? *The Economic Journal* 109:16–24. [arRH]

Binmore, K., Shaked, A. & Sutton, J. (1985) Testing non-cooperative game theory: A preliminary study. *American Economic Review* 75:1178–80. [RS]

Birnbaum, M. & Mellers, B. A. (1983) Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology* 45:792–804. [aRH]

Birnbaum, M. H. & Stegner, S. E. (1979) Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Experimental and Social Psychology* 37:48–74. [EUW]

Blass, T. (1984) Social psychology and personality: Toward a convergence. *Journal of Personality and Social Psychology* 47:1013–27. [RS]

Blume, A. & Ortmann, A. (2000) The effects of costless pre-play communication: Experimental evidence from a game with Pareto-ranked equilibria. (submitted). [rRH]

Bohnet, I., Frey, B. & Huck, S. (2000) More enforcement with less law: On contract enforcement, trust and crowding. KSG Working Paper RWP00–009, Harvard University. http://ksgnotes1.harvard.edu/Research/wpaper.nsf/pubwzAuthor?OpenForm&ExpandView [DJZ]

Bonetti, S. (1998) Experimental economics and deception. *Journal of Economic Psychology* 19:377–95. [aRH]

Bornstein, B. H. (1999) The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior* 23:75–91. [HPD]

Bower, B. (1997) Null science: Psychology's statistical status quo draws fire. *Science News* 151:356–57. [aRH]

Brehmer, B. (1980) In one word: Not from experience. *Acta Psychologica* 45:223–41. [aRH]

(1992) Dynamic decision making: Human control of complex systems. *Acta Psychologica* 81:211–41. [aRH]

(1996) Man as a stabiliser of systems: From static snapshots of judgment processes to dynamic decision making. *Thinking and Reasoning* 2:225–38. [aRH]

Breland, K. & Breland, M. (1961) The misbehavior of organisms. *American Psychologist* 16:681–84. [RK]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Bröder, A. (1998) Deception can be acceptable: Comment on Ortmann and Hertwig. *American Psychologist* 58:805–806. [aRH]
Bruce, D. (1985) The how and why of ecological memory. *Journal of Experimental Psychology: General* 114:78–90. [aRH]
Bruner, J. S. (1957) Going beyond the information given. In: *Contemporary approaches to cognition,* ed. J. S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood & D. Rapaport. Harvard University Press. [TB]
Brunswik, E. (1955) Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62:193–217. [rRH]
Burnham, T., McCabe, K. & Smith, V. L. (2000) Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization* 43:57–73. [rRH]
Butera, F., Mugny, G., Legrenzi, P. & Perez, J. A. (1996) Majority and minority influence, task representation and inductive reasoning. *British Journal of Social Psychology* 35:123–36. [aRH]
Camerer, C. F. (1990) Do markets correct biases in probability judgment? Evidence from market experiments. In: *Advances in behavioral economics,* ed. L. Green & J. Kagel. Ablex. [aRH]
(1995) Individual decision making. In: *The handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [GWH, aRH, DJZ]
(1997) Rules for experimenting in psychology and economics, and why they differ. In: *Understanding strategic interaction. Essays in honor of Reinhard Selten,* ed. W. Albers, W. Güth, P. Hammerstein, B. Molduvanu & E. van Damme. Springer. [arRH]
Camerer, C. F. & Ho, T. (1999) Experience-weighted attraction learning (EWA) in "normal-form" games: A unifying approach. *Econometrica* 67:827–74. [aRH, RK]
Camerer, C. F. & Hogarth, R. M. (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19:7–42. [arRH, RMH, AK]
Camerer, C. F., Johnson, E., Rymon, T. & Sen, S. (1993) Cognition and framing in sequential bargaining for gains and losses. In: *Frontiers of game theory,* ed. K. Binmore, A. Kirman & P. Tani. MIT Press. [rRH]
Camerer, C. F., Loewenstein, G. & Weber, M. (1989) The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy* 97:1232–54. [arRH]
Cameron, J. & Pierce, W. D. (1994) Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research* 64:363–423. [aRH]
Cameron, L. (1999) Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry* 37(1):47–59. [JH]
Campbell, D. T. & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81–105. [FJG-W]
Case, D. & Fantino, E. (1989) Instructions and reinforcement in the observing behavior of adults and children. *Learning and Motivation* 20:373–412. [EF]
Case, D., Fantino, E. & Goodie, A. (1999) Base-rate training without case cues reduces base-rate neglect. *Psychonomic Bulletin and Review* 6:319–27. [EF]
Catania, J., Chitwood, D. D., Gibson, D. R. & Coates, T. J. (1990) Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin* 108:339–62. [DJW]
Chapman, G. B. & Elstein, A. S. (1995) Valuing the future: Temporal discounting of health and money. *Medical Decision Making* 15:373–86. [TR]
Chapman, G. B. & Johnson, E. J. (1995) Preference reversals in monetary and life expectancy evaluations. *Organizational Behavior and Human Decision Processes* 62(3):300–17. [TR]
Charness, G. & Rabin, M. (2000) Social preferences: Some simple tests and a new model. Working paper, Department of Economics, University of California, Berkeley. [rRH]
Chomsky, N. (1975) *Reflections on language.* Random House. [RK]
Christensen, L. (1977) The negative subject: Myth, reality, or a prior experimental experience effect? *Journal of Personality and Social Psychology* 35:392–400. [aRH]
(1988) Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin* 14:664–75. [aRH]
Chu, Y. P. & Chu, R. L. (1990) The subsidence of preference reversals in simplified and marketlike experimental settings: A note. *The American Economic Review* 80:902- 11. [aRH]
Clark, A. (1997) *Being there: Putting brain, body, and world together again.* MIT Press. [DJZ]
Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 69:145–53. [aRH]
(1988) *Statistical power analysis for the behavioral sciences, 2nd edition.* Erlbaum. [aRH]
Connolly, T. (1988) Hedge-clipping, tree-felling, and the management of ambiguity. In: *Managing the challenge of ambiguity and change,* ed. M. B. McCaskey, L. R. Pondy & H. Thomas. Wiley. [aRH]
Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimentation: Design and analysis issues for field settings.* Rand-McNally. [TR]
Cook, T. D., Bean, J. R., Calder, B. J., Frey, R., Krovetz, M. L. & Reisman, S. R. (1970) Demand characteristics and three conceptions of the frequently deceived subject. *Journal of Personality and Social Psychology* 14:185–94. [aRH]
Cosmides, L. & Tooby, J. (1992) Cognitive adaptations for social exchange. In: *The adapted mind: Evolutionary psychology and the generation of culture,* ed. J. H. Barkow, L. Cosmides and J. Tooby. Oxford University Press. [rRH]
(1996) Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58:1–73. [arRH, DSL]
Costa-Gomes, M., Crawford, V. & Broseta, B. (forthcoming) Cognition and behavior in normal-form games: An experimental study. *Econometrica.* [rRH]
Creyer, E. H., Bettman, J. R. & Payne, J. W. (1990) The impact of accuracy and effort feedback and goals on adaptive decision behavior. *Journal of Behavioral Decision Making* 3:1–16. [aRH]
Danziger, K. (1990) *Constructing the subject. Historical origins of psychological research.* Cambridge University Press. [GG, aRH]
Davies, M. F. (1992) Field dependence and hindsight bias: Cognitive restructuring and the generation of reasons. *Journal of Research in Personality* 26:58–74. [arRH]
Davis, D. D. & Holt, C. A. (1993) *Experimental economics.* Princeton University Press. [GWH, arRH]
Dawes, R. M. (1988) *Rational choice in an uncertain world.* Harcourt Brace Jovanovich. [aRH]
(1996) The purpose of experiments: Ecological validity versus comparing hypotheses. *Behavioral and Brain Sciences* 19:20. [aRH]
(1999) Experimental demand, clear incentives, both, or neither? In: *Games and human behavior,* ed. D. V. Budescu, I. Erev & R. Zwick. Erlbaum. [RS]
Deci, E. L., Koestner, R. & Ryan, R. M. (1999a) Meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125:627–68. [aRH]
(1999b) The undermining effect is a reality after all – extrinsic rewards, task interest, and self-determination: Reply to Eisenberger, Pierce, and Cameron (1999) and Lepper, Henderlong, and Gingras (1999). *Psychological Bulletin* 125:692–700. [aRH]
Deci, E. L. & Ryan, R. M. (1987) The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology* 53:1024–37. [HPD]
Dickhaut, J., Hubbard, J. & McCabe, K. (1995) Trust, reciprocity, and interpersonal history: Fool me once, shame on you, fool me twice, shame on me. Working paper, University of Minnesota. [aRH]
Diehl, E. & Sterman, J. D. (1995) Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes* 62:198–215. [aRH]
Diener, E. & Crandall, R. (1978) *Ethics in social and behavioral research.* University of Chicago Press. [aRH]
Dufwenberg, M. & Gneezy, M. (2000) Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30:163–82. [DJZ]
Dufwenberg, M. & Kirchsteiger, G. (2000) A theory of sequential reciprocity. Working paper, Center for Economic Research, Tilburg University. [rRH]
Duhem, P. (1906/1954) *The aim and structure of physical theory,* trans. P. P. Weiner. Princeton University Press. [ASGi]
(1953) Physical theory and experiment. In: *Readings in the philosophy of science,* ed. H. Feigl & M. Brodbeck. Appleton-Century-Crofts. [aRH]
Dulaney, D. E. & Hilton, D. J. (1991) Conversational implicature, conscious representation and the conjunction fallacy. *Social Cognition* 9:85–110. [TB]
Dunwoody, P. T. (2000) The use of base rate information as a function of experienced consistency and utility. Unpublished doctoral dissertation, University of Georgia. [ASGo]
Dyer, D. & Kagel, J. H. (1996) Bidding in common value auctions: How the commercial construction industry corrects for the winners' curse. *Management Science* 42:1463–75. [aRH]
Edwards, W. (1954) The reliability of probability preferences. *American Journal of Psychology* 67:68–95. [rRH]
(1961) Costs and payoffs are instructions. *Psychological Review* 68:275–84. [aRH]
(1962) Dynamic decision theory and probabilistic information processing. *Human Factors* 4:59–73. [aRH]
Eisenberger, R. & Cameron, J. (1996) Detrimental effects of reward: Reality or myth? *American Psychologist* 51:1153–66. [aRH]
Eisenberger, R., Pierce, W. D. & Cameron, J. (1999) Effects of reward on intrinsic motivation – negative, neutral, and positive: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin* 125:677–91. [aRH]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Epley, N. & Huff, C. (1998) Suspicion, affective response, and educational benefit as a result of deception in psychology research. *Personality and Social Psychology Bulletin* 24:759–68. [aRH]
Epstein, Y. M., Suedfeld, P. & Silverstein, S. J. (1973) The experimental contract: Subjects' expectations of and reactions to some behaviors of experimenters. *American Psychologist* 28:212–21. [aRH]
Erev, I., Wallsten, T. S. & Budescu, D. V. (1994) Simultaneous over- and under-confidence: The role of error in judgment processes. *Psychological Review* 101:519–27. [rRH, EUW]
Evans, J. St. B. T. & Bradshaw, H. (1986) Estimating sample-size requirements in research design: A study of intuitive statistical judgment. *Current Psychological Research and Reviews* 5:10–19. [aRH]
Evans, J. St. B. T., Newstead, S. E. & Byrne, R. M. J. (1993) *Human reasoning: The psychology of deduction.* Erlbaum. [aRH]
Evans, J. St. B. T., Over, D. E. & Manktelow, K. I. (1993) Reasoning, decision making and rationality. *Cognition* 49:165–87. [aRH]
Falk, A. & Fischbacher, U. (1999) A theory of reciprocity. Working paper, Institute for Empirical Research in Economics, University of Zurich. [rRH]
Falkenhainer, D., Forbus, K. D. & Gentner, D. (1989) The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41(1):1–63. [ABM]
Fantino, E. (1966) Immediate reward followed by extinction vs. later reward without extinction. *Psychonomic Science* 6:233–34. [EF]
(1998) Judgment and decision making: Behavioral approaches. *The Behavior Analyst* 21:203–18. [EF]
Fehr, E. & Falk, A. (1999) Wage rigidities in a competitive incomplete contract market. An experimental investigation. *Journal of Political Economy* 107:106–34. [DJZ]
Fehr, E. & Gachter, S. (forthcoming) Cooperation and punishment in public goods experiments. *American Economic Review.* [DJZ]
Fehr, E., Kirchsteiger, G. & Riedl, A. (1993) Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108:439–59. [DJZ]
Festinger, L. (1957) *A theory of cognitive dissonance.* Stanford University Press. [MVV]
Festinger, L. & Carlsmith, J. M. (1959) Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology* 58:203–210. [DJH]
Fillenbaum, S. (1966) Prior deception and subsequent experimental performance: The "faithful" subject. *Journal of Personality and Social Psychology* 4:532–37. [aRH]
Fischbein, E. & Schnarch, D. (1997) The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education* 28:96–105. [M-PL]
Fischhoff, B., Slovic, P. & Lichtenstein, S. (1979) Subjective sensitivity analysis. *Organizational Behavior and Human Performance* 23:339–59. [aRH]
Fisher, C. B. & Fyrberg, D. (1994) Participant partners: College students weigh the costs and benefits of deceptive research. *American Psychologist* 49:417–27. [aRH]
Fisher, R. A. (1956) *Statistical methods and scientific inference.* Hafner. [CFB]
Frey, B. S. (1997) *Not just for the money: An economic theory of personal motivation.* Edward Elgar. [aRH]
Frey, B. S. & Oberholzer-Gee, F. (1997) The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review* 87:746–55. [DJZ]
Frey, B. S., Oberholzer-Gee, F. & Eichenberger, R. (1996) The old lady visits your backyard: A tale of morals and markets. *Journal of Political Economy* 104:1297–313. [DJZ]
Fudenberg, D. & Levine, D. (1998) *The theory of learning in games.* MIT Press. [ASGi]
Garcia, J. & Koelling, R. A. (1966) Relation of cue to consequence in avoidance learning. *Psychonomic Science* 4:123–24. [RK]
Garnham, A. & Oakhill, J. (1994) *Thinking and reasoning.* Blackwell. [aRH]
Gazzaniga, M., ed. (1995) *The cognitive neuroscience.* MIT Press. [DSL]
Geller, D. M. (1978) Involvement in role-playing simulations: A demonstration with studies on obedience. *Journal of Personality and Social Psychology* 36:219–35. [aRH]
Gentner, D. & Markman, A. B. (1997) Structural alignment in analogy and similarity. *American Psychologist* 52(1):45–56. [ABM]
Gigerenzer, G. (1991a) How to make cognitive illusions disappear: Beyond heuristics and biases. In: *European review of social psychology, vol. 2,* ed. W. Stroebe & M. Hewstone. Wiley. [rRH]
(1991b) From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review* 98:254–67. [rRH]
(1996) On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review* 103:592–96. [arRH]
Gigerenzer, G., Hell, W. & Blank, H. (1988) Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance* 14:513–25. [CRMM]
Gigerenzer, G. & Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102:684–704. [GG, arRH]
(1999) Overcoming difficulties in Bayesian reasoning: A reply to Lewis & Keren and Mellers & McGraw. *Psychological Review* 106:425–30. [GG]
Gigerenzer, G. & Hug, K. (1992) Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition* 43:127–71. [GG, aRH]
Gigerenzer, G. & Murray, D. J. (1987) *Cognition as intuitive statistics.* Erlbaum. [aRH]
Gigerenzer, G., Todd, P. M. & ABC Research Group (2000) *Simple heuristics that make us smart.* Oxford University Press. [GWH]
Gil-White, F. J. (2001) Ultimatum game with an ethnicity manipulation: Results from Bulgan Cum, Mongolia. In: *Cooperation, punishment, and reciprocity: Experiments in 16 small-scale societies,* ed. J. Henrich, R. Boyd & H. Gintis. (forthcoming) [FJG-W]
Gintis, H. (2000) *Game theory evolving.* Princeton University Press. [HG]
Glanzer, M. & Bowles, N. (1976) Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory* 2:21–31. [CRMM]
Glass, G. V., McGaw, B. & Smith, M. L. (1981) *Meta-analysis in social research.* Sage. [aRH]
Gneezy, U. & Rustichini, A. (forthcoming) A fine is a price. *Journal of Legal Studies.* [DJZ]
Gode, D. K. & Sunder, S. (1993) Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy* 101:119–37. [DJZ]
Goeree, J. K. & Holt, C. A. (2000) A model of noisy introspection. Working paper, Economics Department, University of Virginia. [rRH]
(in press) Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review.* [arRH]
Goldstein, W. M. & Weber, E. U. (1997) Content and discontent: Indications and implications of domain specificity in preferential decision making. In: *Research on judgment and decision making: Currents, connections, and controversies,* ed. W. M. Goldstein & R. M. Hogarth. Cambridge University Press. [aRH]
Goldstone, R. L., Medin, D. L. & Gentner, D. (1991) Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology* 23:222–62. [ABM]
Goltz, S. M. (1993) Examining the joint roles of responsibility and reinforcement history in recommitment. *Decision Sciences* 24:977–94. [EF]
(1999) Can't stop on a dime: The roles of matching and momentum in persistence of commitment. *Journal of Organizational Behavior Management* 19:37–63. [EF]
Goodie, A. S. & Fantino, E. (1995) An experientially derived base-rate error in humans. *Psychological Science* 6:101–106. [EF, ASGo]
(1996) Learning to commit or avoid the base-rate error. *Nature* 380:247–49. [EF, ASGo]
(1999a) What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making* 12:307–35. [EF, ASGo]
(1999b) Base rates versus sample accuracy: Competition for control in human matching to sample. *Journal of the Experimental Analysis of Behavior* 71:155–69. [EF, ASGo]
Goodie, A. S., Ortmann, A., Davis, J. N., Bullock, S. & Werner, G. M. (1999) Demons versus heuristics in artificial intelligence, behavioral ecology, and economics. In: *Simple heuristics that make us smart,* ed. G. Gigerenzer, P. M. Todd & the ABC Research Group. Oxford University Press. [rRH]
Greenwald, A. G., Pratkanis, A. R., Lieppe, M. R. & Baumgardner, M. H. (1986) Under what conditions does theory obstruct research progress? *Psychological Review* 93(2):216–29. [RS]
Greenwood, J. D. (1983) Role-playing as an experimental strategy in social psychology. *European Journal of Social Psychology* 13:235–54. [aRH]
Grether, D. M. (1978) Recent psychological studies of behavior under uncertainty. *American Economic Review* 68:70–74. [rRH]
(1980) Bayes rule as a description model: The representativeness heuristic. *Quarterly Journal of Economics* 95:537–57. [arRH]
(1992) Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior and Organization* 17:31–57. [arRH]
Grether, D. M. & Plott, C. R. (1979) Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69:623–38. [FG, arRH]
Grice, G. R. (1966) Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin* 66:488–99. [HPD]
Griggs, R. A. & Cox, J. R. (1982) The elusive thematic materials effect in the Wason's selection task. *British Journal of Psychology* 73:407–20. [aRH]
Guala, F. (1999) The problem of external validity (or 'parallelism') in experimental economics. *Social Science Information* 38:555–73. [FG]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

(2000) From the laboratory to the outside world: Three routes towards external validity. University of Exeter Working Paper. [FG]
Guidagni, P. M. & Little, J. D. C. (1983) A logit model of brand choice calibrated on scanner data. *Marketing Science* 2:203–38. [ABM]
Güth, W., Schmittberger, R. & Schwarz, B. (1982) An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3:367–88. [HG]
Güth, W. & Tietz, P. (1986) Auctioning ultimatum bargaining positions: How to act if rational decisions are unacceptable. In: *Current issues in West German decision research,* ed. R. W. Scholz. P. Lange. [RS]
Haberstroh, S., Betsch, T. & Aarts, H. (2000) When guessing is better than thinking: Multiple bases for frequency judgments. Submitted for publication. [TB]
Hammond, K. R., Stewart, T. R., Brehmer, B. & Steinman, D. O. (1975) Social judgment theory. In: *Human judgment and decision processes,* ed. M. F. Kaplan & S. Schwartz. Academic Press. [aRH]
Harless, D. W. & Camerer, C. F. (1994) The predictive utility of generalized expected utility theories. *Econometrica* 62:1251–89. [aRH]
Harrison, G. W. (1989) Theory and misbehavior of first-price auctions. *American Economic Review* 79:749–62. [GWH, aRH]
(1992) Theory and misbehavior of first-price auctions: Reply. *American Economic Review* 82:1426–43. [GWH, aRH]
(1994) Expected utility theory and the experimentalists. *Empirical Economics* 19:223- 53. [arRH]
(1999) *Experimental economics and contingent valuation.* Working Paper 96–10, Division of Research, The Darla Moore School of Business, University of South Carolina (http://theweb.badm.sc.edu/glenn/eecv.pdf). [arRH]
Harrison, G. W. & Rutstroem, E. E. (in press) Experimental evidence on the existence of hypothetical bias in value elicitation methods. In: *Handbook of experimental economics results,* ed. C. R. Plott & V. L. Smith. Elsevier. [aRH]
Hartl, J. A. & Fantino, E. (1996) Choice as a function of reinforcement ratios in delayed matching to sample. *Journal of the Experimental Analysis of Behavior* 66:11–27. [EF]
Hedges, L. V. & Olkin, I. (1985) *Statistical methods for meta-analysis.* Academic Press. [aRH]
Hell, W., Gigerenzer, G., Gauggel, S., Mall, M. & Müller, M. (1988) Hindsight bias: An interaction of automatic and motivational factors? *Memory and Cognition* 16:533- 38. [aRH]
Henrich, J. (2000) Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* 90(4):973–79. [JH, rRH]
Henrich, J. & Gil-White, F. (2001) The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior* 22:1–32. [JH]
Hertwig, R. & Gigerenzer, G. (1999) The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making* 12:275–305. [GG, arRH]
Hertwig, R. & Ortmann, A. (2000) Does deception destroy experimental control? A review of the evidence. (submitted). [rRH]
(in press) Economists' and psychologists' experimental practices: How they differ, why they differ, and how they could converge. In: *Economics and psychology,* ed. I. Brocas & J. D. Carillo. Oxford University Press. [rRH]
Hertwig, R. & Todd, P. M. (2000) Biases to the left, fallacies to the right: Stuck in the middle with null hypothesis significance testing. Commentary on Krueger on social-bias. *Psycoloquy* 11(28). http://www.cogsci.soton.ac.uk/psyc- bin/newpsy?11.028. [rRH]
(in press) More is not always better: The benefits of cognitive limits. In: *Reasoning and decision making: A handbook,* ed. D. Hardman & L. Macchi. Wiley. [rRH]
Hey, J. D. (1991) *Experiments in economics.* Blackwell. [arRH]
(1998) Experimental economics and deception: A comment. *Journal of Economic Psychology* 19:397–401. [aRH]
Hilton, D. (1995) The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin* 118:248–71. [aRH]
Hinsz, V. B., Tindale, R. S., Nagao, D. H., Davis, J. H. & Robertson, B. A. (1988) The influence of the accuracy of individuating information on the use of base rate information in probability judgment. *Journal of Experimental Social Psychology* 24:127–45. [aRH]
Hoffman, E., McCabe, K. A., Shachat, K. & Smith, V. (1994) Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7:346–80. [RS]
Hoffman, E., McCabe, K. A. & Smith, V. L. (1996) Social distance and other-regarding behavior in dictator games. *American Economic Review* 86:653–60. [aRH]
(2000) The impact of exchange context on the activation of equity in ultimatum games. *Experimental Economics* 3:5–9. [rRH]
Hoffrage, U. & Hertwig, R. (1999) Hindsight bias: A price worth paying for fast and frugal memory. In: *Simple heuristics that make us smart,* ed. G. Gigerenzer, P. M. Todd & the ABC Research Group. Oxford University Press. [aRH]
Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000) Communicating statistical information. *Science* 290:2261–62. [rRH]
Hogarth, R. M. (1981) Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin* 90:197–217. [aRH, RMH]
Hogarth, R. M., Gibbs, B. J., McKenzie, C. R. M. & Marquis, M. A. (1991) Learning from feedback: Exactness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17:734–52. [aRH]
Hogarth, R. M. & Reder, M. W. (1987) Introduction: Perspectives from economics and psychology. In: *Rational choice: The contrast between economics and psychology,* ed. R. M. Hogarth & M. W. Reder. University of Chicago Press. [aRH]
Holt, C. A. (1995) Industrial organization: A survey of laboratory research. In: *The handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [GWH]
Holt, C. A. & Laury, S. K. (2000) Risk aversion and incentive effects. Working Paper, Department of Economics, University of Virginia. http://www.gsu.edu/~ecoskl [CAH]
Holyoak, K. J. & Thagard, P. (1995) *Mental leaps: Analogy in creative thought.* MIT Press. [M-PL]
Huettel, S. A. & Lockhead, G. R. (2000) Psychologically rational choice: Selection between alternatives in a multiple-equilibrium game. *Cognitive Systems Research* 1(3):143–60. [SAH]
Hulland, J. S. & Kleinmuntz, D. N. (1994) Factors influencing the use of internal summary evaluations versus external information in choice. *Journal of Behavioral Decision Making* 7:79–102. [aRH]
Hummel, J. E. & Holyoak, K. J. (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104(3):427–66. [ABM]
Hunter, J. E. & Schmidt, F. L. (1990) *Methods of meta-analysis: Correcting error and bias in research findings.* Sage. [aRH]
Igou, E. R., Bless, H. & Schenk, W. (1999) Stärkere Framing Effekte durch mehr Nachdenken? Einflüße der Bearbeitungszeit auf Lösungen des "Asian-disease"-Problems [Does more thought enhance framing effects? The impact of extensive processing on solutions to the "Asian disease" problem]. *Sonderforschungsbereich* 504, University of Mannheim, Working Paper 99–11. [TB]
Irwin, J. R., McClelland, G. H. & Schulze, W. D. (1992) Hypothetical and real consequences in experimental auctions for insurance against low-probability risks. *Journal of Behavioral Decision Making* 5:107–16. [aRH]
Jacoby, J., Johar, G. V. & Morrin, M. (1998) Consumer behavior: A quadrennium. *Annual Review of Psychology* 49:319–44. [ABM]
Jamal, K. & Sunder, S. (1991) Money vs. gaming: Effects of salient monetary payments in double oral auctions. *Organizational Behavior and Human Decision Processes* 49:151–66. [aRH]
Jaynes, J. (1976) *The origin of consciousness in the breakdown of the bicameral mind.* Addison-Wesley. [TB]
Jenkins, G. D., Jr., Mitra, A., Gupta, N. & Shaw, J. D. (1998) Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology* 83:777–87. [aRH]
Johar, G. V. (1995) Consumer involvement and deception from implied advertising claims. *Journal of Marketing Research* 32:267–79. [ABM]
Juslin, P., Winman, A. & Olsson, H. (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review* 107:384–96. [rRH]
Kagel, J. H. & Roth, A. E., eds. (1995) *Handbook of experimental economics.* Princeton University Press. [AER]
Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1987) Fairness and the assumptions of economics. In: *Rational choice: The contrast between economics and psychology,* ed. R. M. Hogarth & M. W. Reder. University of Chicago Press. [TB]
Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press. [aRH, RMH]
Kahneman, D. & Tversky, A. (1973) On the psychology of prediction. *Psychological Review* 80:237–51. [arRH, CRMM]
(1979) Prospect theory: An analysis of choice under risk. *Econometrica* 47:263–91. [CAH]
(1996) On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review* 103:582–91. [arRH]
Kahneman, D. & Tversky, A., eds. (2000) *Choices, values, and frames.* Cambridge University Press. [RMH]
Kelman, H. C. (1967) Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin* 67:1–11. [aRH]
Keren, G. & Wegenaar, W. A. (1987) Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:387–91. [rRH]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Kimmel, A. J. (1996) *Ethical issues in behavioral research.* Blackwell. [aRH]
(1998) In defense of deception. *American Psychologist* 53:803–805. [aRH]
Klayman, J. (1988) On the how and why (not) of learning from outcomes. In: *Human judgment: The SJT view,* ed. B. Brehmer & C. R. B. Joyce. North-Holland. [aRH]
Koehler, J. J. (1996) The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences* 19:1–53. [aRH]
Kohn, A. (1996) By all available means: Cameron and Pierce's defense of extrinsic motivators. *Review of Educational Research* 66:1–4. [aRH]
Koriat, A. & Goldsmith, M. (1994) Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General* 123:297–315. [aRH]
(1996) Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences* 19:167- 228. [aRH]
Kreps, D. M. (1990) *A course in microeconomic theory.* Princeton University Press. [aRH]
Kroll, Y. & Levy, H. (1992) Further tests of the separation theorem and the capital asset pricing model. *American Economic Review* 82(3):664–70. [JH]
Krueger, J. (1998) The bet on bias: A foregone conclusion? *Psycoloquy* 9(46). ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/psyc.98.9.46.social-bias.1.krueger ; http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46 [rRH]
Krupat, E. (1977) A re-assessment of role playing as a technique in social psychology. *Personality and Social Psychology Bulletin* 3:498–504. [aRH]
Krupat, E. & Garonzik, R. (1994) Subjects' expectations and the search for alternatives to deception in social psychology. *British Journal of Social Psychology* 33:211–22. [aRH]
Kühberger, A., Schulte-Mecklenbeck, M. & Perner, J. (2000) Framing decisions: Hypothetical and real. Submitted for publication. [AK]
Kuhn, T. S. (1962) *The structure of scientific revolutions.* University of Chicago Press. [FG]
Laury, S. K. & Holt, C. A. (2000) Further reflections on prospect theory. Working Paper, Department of Economics, University of Virginia. (http://www.people.virginia.edu/~cah2k/reflect.pdf). [CAH]
Lecoutre, B., Lecoutre, M.-P. & Poitevineau, J. (2001) Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review.* (forthcoming). [M-PL]
Lecoutre, M.-P. (2000) And . . . what about the researcher's point of view? In: *New ways in statistical methodology: From significance significance tests to Bayesian inference, 2nd edition,* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux. Peter Lang. [M-PL]
Ledyard, J. O. (1995) Public goods: A survey of experimental research. In: *The handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [aRH]
Lepper, M. R., Greene, D. & Nisbett, R. E. (1973) Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology* 28:129–37. [aRH]
Lepper, M. R., Henderlong, J. & Gingras, I. (1999) Understanding the effects of extrinsic rewards on intrinsic motivation – uses and abuses of meta-analysis: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin* 125:669–76. [aRH]
Lepper, M. R., Keavney, M. & Drake, M. (1996) Intrinsic motivation and extrinsic rewards: A commentary on Cameron and Pierce's meta-analysis. *Review of Educational Research* 66:5–32. [aRH]
Leven, S. J. & Levine, D. S. (1995) Parts and wholes: Connectionist dynamics of economic behavior in context. In: *Chaos theory in psychology,* ed. F. D. Abraham & A. Gilger. Greenwood. [DSL]
(1996) Multiattribute decision making in context: A dynamic neural network methodology. *Cognitive Science* 20:271–99. [DSL]
Levin, I. P., Chapman, D. P. & Johnson, R. D. (1988) Confidence in judgments based on incomplete information: An investigation using both hypothetical and real gambles. *Journal of Behavioral Decision Making* 1:29–41. [aRH]
Levine, D. S. (2000) *Introduction to neural and cognitive modeling, 2nd edition.* Erlbaum. [DSL]
Lewin, K. (1935) *A dynamic theory of personality.* McGraw-Hill. [RS]
(1947) Group decision and social change. In: *Readings in social psychology,* ed. T. M. Newcomb & E. L. Hartley. Holt. [RS]
Lewin, S. B. (1996) Economics and psychology: Lessons for our own day from the early twentieth century. *Journal of Economic Literature* 34:1293–323. [DJH]
Lichtenstein, S. & Slovic, P. (1971) Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89:46–55. [aRH]
Lindeman, S. T., van den Brink, W. P. & Hoogstraten, J. (1988) Effect of feedback on base-rate utilization. *Perceptual and Motor Skills* 67:343–50. [aRH]
Lipsey, R. (1979) *An introduction to positive economics, 5th edition.* Weidenfeld and Nicholson. [aRH]
Lockhead, G. R. (1992) Psychophysical scaling: Judgments of attributes or objects? *Behavioral and Brain Sciences* 15:543–601. [SAH]
Lodge, D. (1995) *Small world.* Penguin Books. [aRH]
Loewenstein, G. (1999) Experimental economics from the vantage-point of behavioural economics. *The Economics Journal* 109:F25-F34. [aRH, AK, DJZ]
Logue, A. (1988) Research on self-control: An integrating framework. *Behavioral and Brain Sciences* 11:665–709. [EF]
Loomes, G. (1999) Some lessons from past experiments and some challenges for the future. *The Economic Journal* 109:35–45. [aRH]
Lopes, L. L. (1987) Procedural debiasing. *Acta Psychologica* 64:167–85. [aRH]
(1991) The rhetoric of irrationality. *Theory and Psychology* 1:65–82. [VLS]
(1994) Psychology and economics: Perspectives on risk, cooperation, and the marketplace. *Annual Review of Psychology* 45:197–227. [aRH]
Lowe, C. F. (1980) Determinants of human operant behavior. In: *Advances in the analysis of behavior, vol. 1,* ed. P. Harzem & M. D. Zeiler. Wiley. [ASGo]
MacCoun, R. J. & Kerr, N. L. (1987) Suspicion in the psychological laboratory: Kelman's prophecy revisited. *American Psychologist* 42:199. [aRH]
Manis, M., Dovalina, I., Avis, N. E. & Cardoze, S. (1980) Base rates can affect individual predictions. *Journal of Personality and Social Psychology* 38:231–48. [aRH]
Martindale, C. (1991) *Cognitive psychology: A neural network approach.* Brooks/Cole. [DSL]
Massy, W. F. & Zemsky, R. (1994) Faculty discretionary time: Departments and the "academic ratchet." *Journal of Higher Education* 65:1–22. [aRH]
Matthews, B. A., Shimoff, E., Catania, A. C. & Sagvolden, T. (1977) Uninstructed human responding: Sensitivity to ratio and interval contingencies. *Journal of the Experimental Analysis of Behavior* 27:453–67. [ASGo]
Maule, J. (1989) Positive and negative decision frames: A verbal protocol analysis of the Asian disease problem of Tversky and Kahneman. In: *Process and structure in human decision making,* ed. H. Montgomery & O. Svenson. Wiley. [TB]
Mayr, E. (1988) *The history of biological thought: Diversity, evolution, and inheritance.* Belknap Press. [MM]
McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. (2000) A functional imaging study of "theory of mind" in two-person reciprocal exchange. Working paper, Economic Science Laboratory, University of Arizona. [rRH, VLS]
McCabe, K., Rassenti, S. & Smith, V. L. (1998) Reciprocity, trust, and payoff privacy in extensive form bargaining. *Games and Economic Behavior* 24:10–24. [ASGi]
McCabe, K., Rigdon, M. L. & Smith, V. L. (2000) Cooperation in single play, two-person extensive form games between anonymously matched players. In: *Advances in experimental business research,* ed. R. Zwick & A. Rapoport. Elsevier. [ASGi]
McDaniel, T. & Starmer, C. (1998) Experimental economics and deception: A comment. *Journal of Economic Psychology* 19:403–409. [aRH]
McKenzie, C. R. M. (1998) Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:771- 92. [CRMM]
(1999) (Non)complementary updating of belief in two hypotheses. *Memory and Cognition* 27:152–65. [CRMM]
McKenzie, C. R. M., Wixted, J. T. & Noelle, D. C. (2000) Do participants believe what experimenters tell them? Further tests of the relation between confidence in yes/no and forced-choice tasks. (submitted). [CRMM]
McKenzie, C. R. M., Wixted, J. T., Noelle, D. C. & Gyurjyan, G. (2001) Relation between confidence in yes/no and forced-choice tasks. *Journal of Experimental Psychology: General* 130:140–55. [CRMM]
Medin, D. L. & Bettger, J. G. (1991) Sensitivity to changes in base-rate information. *American Journal of Psychology* 104:311–32. [aRH]
Medin, D. L. & Edelson, S. M. (1988) Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General* 117:68–85. [aRH]
Medin, D. L., Goldstone, R. L. & Gentner, D. (1993) Respects for similarity. *Psychological Review* 100(2):254–78. [ABM]
Meehl, P. E. (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* 46:806–34. [CFB, aRH]
Mellers, B. A., Berretty, P. M. & Birnbaum, M. H. (1995) Dominance violations in judged prices of two- and three-outcome gambles. *Journal of Behavioral Decision Making* 8:201–16. [aRH]
Mellers, B. A., Hertwig, R. & Kahneman, D. (2001) Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science* 12:269–75. [aRH]
Milgram, S. (1963) Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67:371–78. [HPD]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Miller, D. T. (1999) The norm of self-interest. *American Psychologist* 54:1053–60. [MVV]
Mixon, D. (1972) Instead of deception. *Journal for the Theory of Social Behaviour* 2:145- 77. [aRH]
Nass, C., Moon, Y., Morkes, J., Kim, E.-Y. & Fogg, B. J. (1997) Computers are social actors: A review of current research. In: *Moral and ethical issues in human- computer interaction,* ed. B. Friedman. CSLI Press. [JH]
Nelson, T. E., Biernat, M. R. & Manis, M. (1990) Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology* 59:664–75. [aRH]
Newberry, B. H. (1973) Truth telling in subjects with information about experiments: Who is being deceived? *Journal of Personality and Social Psychology* 25:369–74. [aRH]
Nicks, S. D., Korn, J. H. & Mainieri, T. (1997) The rise and fall of deception in social psychology and personality research, 1921 to 1994. *Ethics and Behavior* 7:69–77. [aRH]
Nisbett, R. E. & Cohen, D. (1996) *Culture of honor: The psychology of violence in the South.* Westview Press. [FJG-W]
Nisbett, R., Peng, K., Choi, I. & Norenzayan, A. (in press) Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review.* [JH]
Offerman, T., Sonnemans, J. & Schram, A. (in press) Expectation formation in step-level public good games. *Economic Inquiry.* [rRH]
Olcina, G. & Urbano, A. (1994) Introspection and equilibrium selection in 2x2 matrix games. *International Journal of Game Theory* 23:183–206. [rRH]
Oliansky, A. (1991) A confederate's perspective on deception. *Ethics and Behavior* 1:253–58. [aRH]
Olson, M. J. & Budescu, D. V. (1997) Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making* 10:117–31. [aRH]
Ong, A. D. & Weiss, D. J. (2000) The impact of anonymity on responses to "sensitive" questions. *Journal of Applied Psychology* 30:1691–708. [DJW]
Ordóñez, L. D., Mellers, B. A., Chang, S. J. & Roberts, J. (1995) Are preference reversals reduced when made explicit? *Journal of Behavioral Decision Making* 8:265–77. [aRH]
Orne, M. T. (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17:776–83. [aRH]
Ortmann, A. & Colander, D. C. (1997) A simple principal-agent experiment for the classroom. *Economic Inquiry* 35:443–50. [aRH]
Ortmann, A. & Gigerenzer, G. (1997) Reasoning in economics and psychology: Why social context matters. *Journal of Institutional and Theoretical Economics* 153(4):700–10. [aRH]
Ortmann, A., Fitzgerald, J. & Boeing, C. (2000) Trust, reciprocity, and social history: A re-examination. *Experimental Economics* 3:81–100. [rRH]
Ortmann, A. & Hertwig, R. (1997) Is deception acceptable? *American Psychologist* 52:746–47. [aRH]
(1998) The question remains: Is deception acceptable? *American Psychologist* 53:806- 807. [aRH]
(2000) The costs of deception? Evidence from psychology. (submitted). [rRH]
Ortmann, A. & Squire, R. C. (2000) A game-theoretic explanation of the administrative lattice in institutions of higher learning. *Journal of Economic Behavior and Organization* 43:377–91. [aRH]
Ortmann, A. & Tichy, L. (1999) Understanding gender effects in the laboratory: Evidence from Prisoner's Dilemma games. *Journal of Economic Behavior and Organization* 39:327–39. [aRH]
Osborne, M. J. & Rubinstein, A. (1990) *Bargaining and markets.* Academic Press. [aRH]
Payne, J. W., Bettman, J. R. & Johnson, E. J. (1992) Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology* 43:87–131. [aRH]
(1993) *The adaptive decision maker.* Cambridge University Press. [ABM, EUW]
Payne, J. W., Bettman, J. R. & Luce, M. F. (1996) When time is money: Decision behavior under opportunity-cost time pressure. *Organizational Behavior and Human Decision Processes* 66:131–52. [aRH]
Pelham, B. W. & Neter, E. (1995) The effect of motivation of judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology* 68:581- 94. [aRH]
Piaget, J. (1996) *Structuralism.* Routledge & Kegan Paul. [DJH]
Pillutla, M. M. & Chen, X.-P. (1999) Social norms and cooperation in social dilemmas: The effects of context and feedback. *Organizational Behavior and Human Decision Processes* 78(2):81–103. [JH]
Plott, C. R. (1979) The application of laboratory experimental methods to public choice. In: *Collective decision making: Applications from public choice theory,* ed. C. S. Russell. Johns Hopkins University Press. [HG]
Premack, D. & Woodruff, G. (1978) Chimpanzee problem-solving: A test for comprehension. *Science* 202:53–35. [RK]
Prendergast, C. (1999) The provision of incentives in firms. *Journal of Economic Literature* 37:7–63. [aRH]
Quine, W. V. (1953a) *From a logical point of view.* Harvard University Press. [aRH]
(1953b) Two dogmas of empiricism. In: *From a logical point of view.* Harvard University Press. [ASGi]
Rabin, M. (1998) Psychology and economics. *Journal of Economic Literature* 36:11–46. [arRH]
Rachlin, H. (1995) Self-control: Beyond commitment. *Behavioral and Brain Sciences* 18:109–59. [EF]
Rapoport, A. & Wallsten, T. S. (1972) Individual decision behavior. *Annual Review of Psychology* 23:131–76. [aRH]
Rassenti, S. (1981) *0–1 Decision problems: Algorithms and applications.* Doctoral dissertation, University of Arizona. [VLS]
Reed, S. J. & Miller, L. C., eds. (1998) *Connectionist models of social reasoning and social behavior.* Erlbaum. [DSL]
Riecken, H. W. (1962) A program for research on experiments in social psychology. In: *Decisions, values and groups, vol. 2,* ed. N. F. Washburne. Pergamon. [aRH]
Ring, K. (1967) Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology* 3:113–23. [aRH]
Robertson, I. (2000) Imitative problem solving: Why transfer of learning often fails to occur. *Instructional Science* 28:263–89. [M-PL]
Rosenthal, R. (1990) How are we doing in soft psychology? *American Psychologist* 45:775–77. [FJG-W, aRH]
Rosenthal, R. & Rosnow, R. L. (1975) *The volunteer subject.* Wiley. [MVV]
(1991) *Essentials of behavioral research: Methods and data analysis, 2nd edition.* McGraw Hill. [aRH]
Rossi, J. S. (1990) Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology* 58:646–56. [aRH]
Roth, A. E. (1993) On the early history of experimental economics. *Journal of the History of Economic Thought* 15:184–209. http://www.economics.harvard.edu/~aroth/history.html [AER]
(1995) Introduction to experimental economics. In: *The handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [ASGo, aRH, DSL]
Roth, A. E. & Erev, I. (1995) Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8:164–212. [aRH]
Roth, A. E., Erev, I., Slonin, R. L. & Barron, G. (2000) Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. Working paper, Harvard University. [IE]
Roth, A. E., Prasniker, V., Okuno-Fujiwara, M. & Zamir, S. (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review* 81(5):1068–95. [JH]
Rubin, Z. (1985) Deceiving ourselves about deception: Comment on Smith and Richardson's "amelioration of deception and harm in psychological research." *Journal of Personality and Social Psychology* 48:252–53. [aRH]
Rubin, Z. & Moore, J. C. Jr. (1971) Assessment of subjects' suspicions. *Journal of Personality and Social Psychology* 17:163–70.
Russell, R. R. & Wilkinson, M. (1979) *Microeconomics: A synthesis of modern and neoclassical theory.* Wiley. [aRH]
Russo, J. E. & Dosher, B. A. (1983) Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9(4):676–96. [ABM]
Sally, D. (1995) Conversation and cooperation in social dilemmas. *Rationality and Society* 7(1):58–92. [HG]
Samuelson, P. A. (1963) Discussion contribution. *American Economic Review* 53(2):231- 36. [rRH]
Samuelson, P. A. & Nordhaus, W. (1985) *Principles of economics, 12th edition.* McGraw-Hill. [rRH]
Saxe, L. (1991) Lying: Thoughts of an applied social psychologist. *American Psychologist* 46:409–15. [DJW]
Schmidt, F. L. (1992) What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47:1173–81. [aRH]
Schmidt, F. L. & Hunter, J. E. (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: *What if there were no significance tests?,* ed. L. L. Harlow, S. A. Mulaik & J. H. Steiger. Erlbaum. [rRH]
Schotter, A. & Merlo, A. (1999) A surprise-quiz view of learning in economic experiments. *Games and Economic Behavior* 28:25–54. [rRH]
Schotter, A., Weiss, A. & Zapater, I. (1996) Fairness and survival in ultimatum and dictatorship games. *Journal of Economic Behavior and Organization* 31:37–56. [aRH]
Schulz, D. P. (1969) The human subject in psychological research. *Psychological Bulletin* 72:214–28. [aRH]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

Schwartz, B. (1982) Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General* 111:23–59. [aRH]
(1989) *Psychology of learning and behavior, 3rd edition.* Norton. [RK]
Sedlmeier, P. & Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309–16. [aRH]
Selten, R. (1998) Aspiration adaptation theory. *Journal of Mathematical Psychology* 42:191–214. [aRH]
Shapira, Z. (2000) Aspiration levels and risk taking: A theoretical model and empirical study on the behavior of government bond traders. Unpublished manuscript, New York University. [DJH]
Sharpe, D., Adair, J. G. & Roese, N. J. (1992) Twenty years of deception research: A decline in subjects' trust? *Personality and Social Psychology Bulletin* 18:585–90. [aRH]
Shefrin, H. (2000) *Beyond greed and fear: Understanding behavioral finance and the psychology of investing.* Harvard Business School Press. [DSL]
Shepard, R. N. (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika* 27(2):125–40. [ABM]
Sieber, J. E., Iannuzzo, R. & Rodriguez, B. (1995) Deception methods in psychology: Have they changed in 23 years? *Ethics and Behavior* 5:67–85. [aRH]
Sieber, J. E. & Saks, M. J. (1989) A census of subject pool characteristics and policies. *American Psychologist* 44:1053–61. [arRH]
Siegel, S. (1956) *Nonparametric statistics for the behavioral sciences.* McGraw Hill. [CAH]
Siegel, S. & Goldstein, D. A. (1959) Decision making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology* 57:37–42. [CAH]
Siegel, S. & Fouraker, L. B. (1960) *Bargaining and group decision making.* McGraw Hill. [CAH]
Silverman, I., Shulman, A. D. & Wiesenthal, D. L. (1970) Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology* 14:203–12. [rRH]
Simon, H. A. (1956) Rational choice and the structure of environments. *Psychological Review* 63:129–38. [rRH]
(1957) *Models of man.* Wiley. [aRH]
(1979) *Models of thought.* Yale University Press. [HPD]
(1990) Invariants of human behavior. *Annual Review of Psychology* 41:1–19. [rRH]
Slovic, P., Griffin, D. & Tversky, A. (1990) Compatibility effects in judgment and choice. In: *Insights in decision making: A tribute to Hillel J. Einhorn,* ed. R. M. Hogarth. University of Chicago Press. [EUW]
Smith, A. (1759/1976) *The theory of moral sentiments.* Oxford University Press. [rRH]
Smith, M. J. (1982) *Evolution and the theory of games.* Cambridge University Press. [rRH]
Smith, S. M. & Levin, I. P. (1996) Need for cognition and choice framing effects. *Journal of Behavioral Decision Making* 9:283–90. [aRH]
Smith, S. S. & Richardson, D. (1983) Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology* 44:1075–82. [aRH]
Smith, V. L. (1976) Experimental economics: Induced value theory. *American Economic Review Proceedings* 66:274–79. [arRH]
(1982) Microeconomic systems as an experimental science. *American Economic Review* 72:923–55. [HG, arRH]
(1991) Rational choice: The contrast between economics and psychology. *Journal of Political Economy* 99:877–97. [aRH]
Smith, V. L. & Szidarovszky, F. (2000) Monetary rewards and decision cost in strategic interactions. [VLS]
Smith, V. L. & Walker, J. M. (1993a) Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31:245–61. [arRH]
(1993b) Rewards, experience and decision costs in first price auctions. *Economic Inquiry* 31:237–45. [aRH]
Snyder, M. & Swann, W. B. (1978) Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Personality and Social Psychology* 36:1202–12. [MVV]
Stahl, D. & Wilson, P. (1995) On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10:218–54. [rRH]
Stang, D. J. (1976) Ineffective deception in conformity research: Some causes and consequences. *European Journal of Social Psychology* 6:353–67. [aRH]
Starmer, C. (1999) Experimental economics: Hard science or wasteful tinkering? *The Economic Journal* 109:5–15. [aRH]
Stiggelbout, A. M., deHaes, J. C. J. M., Kiebert, G. M., Kievit, J. & Leer, J. W. H. (1996) Tradeoffs between quality and quantity of life: Development of the QQ Questionnaire for cancer patient attitude. *Medical Decision Maker* 16:184–92. [TR]
Stolarz-Fantino, S. & Fantino, E. (1990) Cognition and behavior analysis: A review of Rachlin's *Judgment, decision and choice. Journal of the Experimental Analysis of Behavior* 54:317–22. [EF]
Stone, D. N. & Schkade, D. A. (1994) Effects of attribute scales on process and performance in multiattribute choice. *Organizational Behavior and Human Decision Processes* 59:261–87. [aRH]
Stone, D. N. & Ziebart, D. A. (1995) A model of financier incentive effects in decision making. *Organizational Behavior and Human Decision Processes* 61:250–61. [aRH]
Svenson, O. (1996) Decision making and the search for fundamental regularities: What can be learned from a process perspective? *Organizational Behavior and Human Decision Processes* 65(3):252–67. [TR]
Taylor, K. M. & Shepperd, J. A. (1996) Probing suspicion among participants in deception research. *American Psychologist* 51:886–87. [aRH]
Thaler, R. (1987) The psychology of choice and the assumptions of economics. In: *Laboratory experimentation in economics: Six points of view,* ed. A. E. Roth. Cambridge University Press. [aRH, DSL]
Tirole, J. (1988) *The theory of industrial organization.* MIT Press. [aRH]
Todd, P. M. & Gigerenzer, G. (2000) Précis of *Simple heuristics that make us smart. Behavioral and Brain Sciences* 23:727–41. [rRH]
Todd, P. M., Gigerenzer, G. & the ABC Research Group (2000) How can we open up the adaptive toolbox? *Behavioral and Brain Sciences* 23:767–80. [rRH]
Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind,* ed. J. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [RK, DR]
Tversky, A. (1977) Features of similarity. *Psychological Review* 84(4):327–52. [ABM]
Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–31. [DSL]
(1981) The framing of decisions and the rationality of choice. *Science* 211:453–58. [aRH, DSL]
(1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90:293–315. [aRH]
(1987) Rational choice and the framing of decisions. In: *Rational choice: The contrast between economics and psychology,* ed. R. M. Hogarth & M. W. Reder. University of Chicago Press. [arRH]
Tversky, A. & Kahneman, D. (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5:297–323. [CAH]
Van Huyck, J., Battalio, R. C., Mathur, S., Van Huyck, P. & Ortmann, A. (1995) On the origin of conventions: Evidence from symmetric bargaining games. *International Journal of Game Theory* 24:187–212. [rRH]
Van Wallendael, L. R. (1995) Implicit diagnosticity in an information-buying task: How do we use the information that we bring with us to a problem? *Journal of Behavioral Decision Making* 8:245–64. [aRH]
Van Wallendael, L. R. & Guignard, Y. (1992) Diagnosticity, confidence, and the need for information. *Journal of Behavioral Decision Making* 5:25–37. [aRH]
Vinacke, W. E. (1954) Deceiving experimental subjects. *American Psychologist* 9:155. [aRH]
Von Winterfeldt, D. & Edwards, W. (1982) Cost and payoffs in perceptual research. *Psychological Bulletin* 91:609–22. [aRH]
Walker, J. & Smith, V. L. (1993) Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31(2):245–61. [VLS]
Wallsten, T. S. (1972) Conjoint-measurement framework for the study of probabilistic information processing. *Psychological Review* 79:245–60. [arRH]
(1976) Using conjoint-measurement models to investigate a theory about probabilistic information processing. *Journal of Mathematical Psychology* 14:144–85. [arRH]
(1982) Research through deception (letter to the editor). *New York Times Magazine,* October 1982. [aRH]
Wallsten, T. S. & Barton, C. (1982) Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8:361–84. [aRH]
Wason, P. C. (1981) The importance of cognitive illusions. *Behavioral and Brain Sciences* 4:356. [aRH]
Weber, E. U. (1994) From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin* 115:228–42. [EUW]
(1998) From Shakespeare to Spielberg: Some reflections on modes of decision making. Presidential Address, Annual Meeting of the Society for Judgment and Decision Making, Dallas, Texas, November 13, 1998. [EUW]
Weber, E. U. & Hsee, C. K. (1999) Models and mosaics: Investigating cross-cultural differences in risk perception and risk preference. *Psychonomic Bulletin and Review* 6:611–17. [EUW]
Weber, E. U. & Kirsner, B. (1997) Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty* 14:41–61. [EUW]
Weber, S. J. & Cook, T. D. (1972) Subject effects in laboratory research: An

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press

examination of subjects role, demand characteristics, and valid inference. *Psychological Bulletin* 77:273–95. [aRH]

Weibull, J. W. (1995) *Evolutionary game theory.* MIT Press. [rRH]

Wilcox, N. T. (1993) Lottery choice: Incentives, complexity and decision time. *Economic Journal* 103:1397–417. [aRH]

Wilson, T. D. & Schooler, J. W. (1991) Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* 60:181–92. [TB]

Winkler, R. L. & Murphy, A. H. (1973) Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance* 10:252–70. [aRH]

Wixted, J. T. (1992) Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:681–90. [CRMM]

Wundt, W. (1892/1894) *Vorlesungen über die Menschen und Thierseele, 2nd edition.* Leopold Voss. English edition: (1894) *Lectures on human and animal psychology,* trans. J. E. Creighton & E. B. Titchener. Swan Sonnenschein. [rRH]

Wyer, R. S., ed. (1997) *The automaticity of everyday life – advances in social cognition, vol. X.* Erlbaum. [TB]

Yaniv, I. & Schul, Y. (1997) Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making* 10:211–20. [aRH]

Young, P. (1993) The evolution of conventions. *Econometrica* 61:57–84. [rRH]

Zhang, S. & Markman, A. B. (1998) Overcoming the early entrant advantage: The role of alignable and nonalignable differences. *Journal of Marketing Research* 35:413–26. [ABM]

Zizzo, D. J. (2001) Transitive and intransitive preferences in choices between simple and compound lotteries: Experimental evidence and neural network modelling. Oxford University Department of Economics Discussion Paper, No. 57. http://www.economics.ox.ac.uk/Research/Ree/cjf2.pdf [DJZ]

Zizzo, D., Stolarz-Fantino, S., Wen, J. & Fantino, E. (2000) A violation of the monotonicity axiom: Experimental evidence on the conjunction fallacy. *Journal of Economic Behavior and Organization* 41:263–76. [EF, DJZ]

Zwick, R., Erev, I. & Budescu, D. V. (1999) The psychological and economical perspectives on human decisions in social and interactive contexts. In: *Games and human behavior: Essays in honor of Amnon Rapoport,* ed. D. V. Budescu, I. Erev & R. Zwick. Erlbaum. [aRH]

https://doi.org/10.1017/S0140525X01354146 Published online by Cambridge University Press