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-

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