AUSTRALIAN
SCHOOL OF BUSINESS™

THE UNIVERSITY OF NEW SOUTH WALES

The University of New South Wales
Australian School of Business

School of Economics Discussion Paper: 2010/06

"The Way in which an Experiment is Conducted is Unbelievably Important": On the Experimentation Practices of Economists and Psychologists

Andreas Ortmann

School of Economics
Australian School of Business
UNSW Sydney NSW 2052 Australia
http://www.economics.unsw.edu.au

ISSN 1837-1035
ISBN 978 0 7334 2905-7

"The Way in which an Experiment is Conducted is Unbelievably Important": On the Experimentation Practices of Economists and Psychologists[1]

Abstract

To discuss experimental results without discussing how they came about makes sense when the results are robust to the way experiments are conducted. Experimental results, however, are – arguably more often than not – sensitive to numerous design and implementation characteristics such as the use of financial incentives, deception, and the way information is presented. To the extent that economists and psychologists have different experimental practices, this claim is of obvious practical and interpretative relevance. In light of the empirical results summarized below, it seems warranted to say that it does not make sense to report experimental results without reporting the design and implementation choices that were made.[2]

Keywords: Duhem-Quine problem, experimental design, experimental implementation, financial incentives, deception.

Andreas Ortmann
Economics Department
Australian School of Business
University of New South Wales
Australia – Sydney
aortmann@yahoo.com

[1] This chapter is based on a presentation at the *25. Hamburger Symposion zur Methodologie der Sozialpsychologie*, Hamburg, January 16, 2009; it draws significantly on joint work with Ralph Hertwig, a professor of cognitive and decision sciences at the Institute for Psychology, University of Basel, Switzerland. I finalized the manuscript during a delightful and productive visit at the Center for Economic Studies in Munich, Germany, in the fall of 2009.

[2] It is therefore good practice (e.g., Camerer, 2003) to discuss design and implementation details and it strikes me as outright silly to "emphasize what psychologists and experimental economists have learned about people, rather than *how* they have learned it." (Rabin, 1998, p. 12).

A surprising fact?

It is a fact, and one that surprised me when I encountered it more than a decade ago, that two related disciplines, economics and corresponding areas in psychology (in particular behavioral decision making), have very different conceptions of what constitutes good experimentation regarding key issues such as the use of scripts, repeated trials, financial incentives, deception, abstract lab environs, and so on.

According to Hertwig and Ortmann (2001), economists bring to experiments a precisely defined "script" for subjects to enact, often use repeated experimental trials to allows subjects to learn about the task and the environment, pay subjects on the basis of clearly defined performance criteria (which are informed by models that assume the maximization of some function), and virtually never deceive their subjects. In contrast, psychologists typically do not provide scripts, do not use repeated experimental trials, pay a flat fee or grant course credit, and do use deception.[3]

Hertwig and Ortmann (2001) argue that economists reduce uncertainty by defining the demand characteristics of the situation as that of performance (script, financial incentives), allowing participants to gain experience with the situation (repetition), and reducing second guessing (no deception). They also argue that in contrast, psychologists leave room for uncertainty by not clearly defining the demand characteristics of the situation (no script, no payment), interpreting people's one-off response as indicative of their general competence (no repetition), and inviting second-guessing. It is not without irony that psychologists are much more laissez-faire than economists.

I emphasize that these statements refer to those areas where economists and psychologists pursue similar topics and questions: behavioral decision making and related areas in social and cognitive psychology such as social cognition, problem solving, and reasoning. These statements do not necessarily apply to research practices in sensation and perception, biological psychology, psycho-physics, neuro-psychology, learning and related fields (behaviorism).

Why these differences? (And why are standards in psychology relatively laissez-faire while in economics standards are rigorously enforced?)

Historically, experimental economists were the new kids on the block who had to prove themselves to a skeptical crowd, according to Hertwig and Ortmann (2001). (And skeptical that crowd was: recall the Wallis-Friedman critique, or read Smith (2008) to get an idea how skeptical!) One obvious way to address the skepticism was

[3] Hertwig and Ortmann (2001) grounded this assessment on empirical exercises. For example, they re-examined the use of repeated experimental trials by going through an authoritative review of Bayesian reasoning studies (Koehler, 1996), they quantified the use of financial incentives by going through all the publications during the years 1988–1997 in the *Journal of Behavioral Decision Making*, and they quantified the use of deception likewise through sampling techniques that guarded to some extent against biased samples. For details, see Hertwig and Ortmann (2001), Ortmann and Hertwig (2002), and Hertwig and Ortmann (2008a, 2008b).

to pre-empt it. Thus, future objections along the lines of the Wallis-Friedman critique were addressed by making decisions matter financially. Apart from there being a good rationale for the use of financial incentives, it also helped that all the way through the 50's, 60's, 70's and 80's experimental economists were working in about half a dozen places (e.g., Purdue, Arizona, Caltech, Frankfurt, Texas A & M), thus facilitating the codification of methodological practices (e.g., Smith, 1976, 1982). The resultant canon of good experimentation has only recently been questioned (e.g., Ortmann & Gigerenzer, 1997; Hertwig & Ortmann, 2001; Harrison & List, 2004; List, 2006; Levitt & List, 2007), with much of the questioning focused on the abstractness of the lab environs; the *semantic* and *pragmatic* ambiguity of natural language; experimenters' violations of conversational norms; questions about information presentation; the control of subject characteristics such as gender, social status, and cognitive abilities; and last but not least the question of external validity.

Do these differences matter?

> "Experimental results always present a joint test of the theory (however well articulated, formally) that motivated the test, and all the things you had to do to implement the test." (Smith 2002, p. 98)
> "The way in which an experiment is conducted is unbelievably important." (Camerer, 2003, p. 34)

While differences in experimentation may be a surprising fact, the relevant question is whether these differences matter. Camerer (2003), in his eminently readable book on behavioral game theory, from which parts of the title of this chapter have been culled, leaves no doubt about it. The reason is known as the Duhem-Quine problem: each test of a theory is a joint test of the theory and the way in which the experiment is conducted. The way an experiment is conducted requires various judgment calls, or "auxiliary hypotheses" (Smith, 2002, p. 98) about the effects of design and implementation decisions such as financial incentives, information presentation, and subject characteristics. To the extent that each of these details can affect the experimental results, the way an experiment is conducted can indeed be extremely important. As it turns out, whether one uses financial incentives, how one presents information, and whether one controls for subject characteristics can all be potentially important determinants of experimental outcomes, as can be the use of deception. Below I shall focus on the first three.

The case of financial incentives

Financial incentives are performance-based participant payments (not necessarily of a monetary kind); they obviously require the existence of a performance standard.

The use of financial incentives in economics is pervasive: for example, Camerer and Hogarth (1999) find that in all experimental studies published in the *American Economic Review* in 1970–1997, subjects were paid according to performance. Following up on their study I find that the same is true for that journal through 2008. (As a matter of fact, the same seems to hold true for other top journals except in those rare cases where the payment mode has been made an explicit focus of the study (e.g., Gneezy & Rustichini, 2000; see also Rydval & Ortmann, 2004).

In contrast, the use of financial incentives in psychology (especially in areas such as judgment and decision making) is not typical operating procedure. Hertwig and Ortmann (2001), for example, report that subjects were paid according to performance in only 26% of 186 experimental studies published in the *Journal of Behavioral Decision Making* during the ten-year period 1988–1997. They stress that this sample identified an upper bound of the use of financial incentives in psychology since the journal publishes articles by psychologists, economists, and management scientists and has experimental economists on the editorial board and as frequent contributors. Indeed, Hertwig and Ortmann (2001) find that financial incentives were used in only 3 out of 106 Bayesian reasoning studies referenced in Koehler (1996) and published in various social psychology, cognitive psychology, and judgment and decision-making journals.

Do financial incentives matter?

Psychologists, for the most part, claim that financial incentives do not matter. Experimental findings allegedly provide little support for the view that the "observed failures of rational models are attributable to the cost of thinking and thus will be eliminated by proper incentives" (Tversky & Kahneman, 1987, p. 90; see also Smith, 1991)

Economists are convinced that financial incentives do matter: Smith and Walker (1993) survey 31 studies and find that financial incentives, while not guaranteeing optimal decisions, in many cases bring decisions closer to the predictions of the normative model; they also reduce data variability. Camerer and Hogarth (1999) analyze 74 studies of "judgments and decisions", "games and markets," and "individual choice" and find that, across all three research domains, in 45% of the studies financial incentives did not make a difference, in 40% they had a positive effect, and in 15% the effects were negative. Re-analyzing these data, Hertwig and Ortmann (2003) find that the positive effect is particularly extensive for experiments in judgment and decision making but less so for market experiments.

Both studies (Smith & Walker, 1993, and Camerer & Hogarth, 1999) are opportunistic samples in that the authors analyzed what they came across in a process that is not clearly documented. Such a procedure invites various selection biases. Hertwig and Ortmann (2001) look at a non-opportunistic sample: the relevant subset of 10 studies (that either compared payment or nonpayment conditions or different payment schemes) in the *Journal of Behavioral Decision Making* sample used to study the use of financial incentives. The authors find that in a clear majority of cases where financial incentives made a difference, they improved participants' performance (and, in fact, did so in exactly the way Smith and Walker, 1993, claim they do: by bringing decisions closer to the predictions of normative models and by reducing error variance substantially). In two cases (of which one was compromised by methodological problems) they led to impaired performance, and in a couple of cases they did not affect performance.

While, thus, the evidence is somewhat divergent on the extent and intensity of the effects of financial incentives, it seems fair to say that there is agreement on at least the following claims:

- Financial incentives matter more in some areas than in others.

- For judgment and decision making, financial incentives seem to matter more often than not.
- The effects, for judgment and decision making and elsewhere, seem to be two-fold:
  - financial incentives move the data closer to the game/decision-theoretic prediction.
  - financial incentives reduce the variability of the data.

Hertwig and Ortmann (2001) argue that this suggests as a basic policy prescription "Do-it-both-ways! (Do-it-n-ways!)" when in doubt (or, if you get very surprising results).

The Hertwig and Ortmann (2001) target article drew more than 30 commentaries and also produced some follow-up studies (Holt & Laury, 2002; Parco, Rapoport & Stein, 2002; Rydval & Ortmann, 2004), which, since they illustrate the potentially stark effects of financial incentives well, I will briefly discuss next.

Holt and Laury (2002). Elaborating on their commentary to Hertwig and Ortmann (2001), these authors study the empirical issues of how risk aversion depends on the size of the financial incentives (stakes). The key assessment instrument they used (now generally known as the Holt and Laury instrument) is two lotteries, one relatively safe (i.e., the two prospects being relatively close together) and the other more risky (i.e., the possible outcomes, or prospects, being wider apart). These lotteries can be seen in Table 1 below (which is reproduced from Holt & Laury, 2002). Specifically, the prospects for lottery ("Option") A were \$1.60 and \$2.00 and those for lottery ("Option") B were \$0.10 and \$3.85. As probability weights are shifted increasingly to the higher prospects in these options, risk-neutral subjects should switch from selecting Option A for the first four rows and Option B from row 5 on. Risk-averse subjects ought to switch in later rows dependent on their degree of risk aversion. In the last row even the most risk-averse subjects should obviously switch.

TABLE 1—THE TEN PAIRED LOTTERY-CHOICE DECISIONS WITH LOW PAYOFFS

Option A	Option B	Expected payoff difference
1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10	\$1.17
2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10	\$0.83
3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10	\$0.50
4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10	\$0.16
5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10	−\$0.18
6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10	−\$0.51
7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10	−\$0.85
8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10	−\$1.18
9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10	−\$1.52
10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10	−\$1.85

[Table 1 reproduced from Holt and Laury (2002), p. 1645]

The empirical question that Holt and Laury (2002) were interested in was: Do people switch at the same point when the stakes are increased (i.e., when the outcomes were multiplied by factors of 20, 50, and 90), and whether it matters if this scaling of the stakes was done hypothetically or indeed for real.

The figures below (which have been reproduced from Holt & Laury, 2002), present the key results of this investigation:

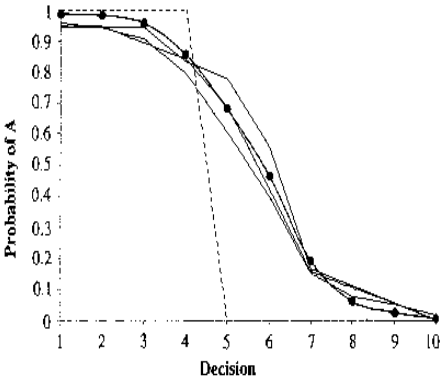

FIGURE 1. PROPORTION OF SAFE CHOICES IN EACH DECISION: DATA AVERAGES AND PREDICTIONS

Note: Data averages for low real payoffs [solid line with dots], 20x, 50x, and 90x hypothetical payoffs [thin lines], and risk-neutral prediction [dashed line].

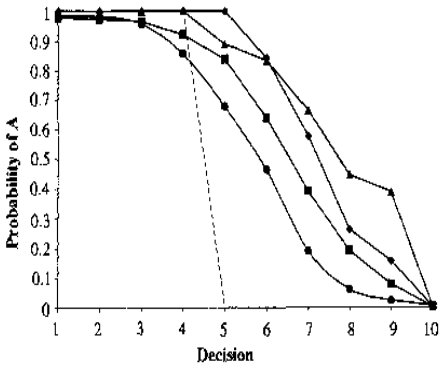

FIGURE 2. PROPORTION OF SAFE CHOICES IN EACH DECISION: DATA AVERAGES AND PREDICTIONS

Note: Data averages for low real payoffs [solid line with dots], 20x real [squares], 50x real [diamonds], 90x real payoffs [triangles], and risk-neutral prediction [dashed line].

[Figures 1 and 2 reproduced from Holt and Laury (2002), pp. 1648, 1649]

Figure 1 demonstrates that the distribution of switching choices remains essentially the same for hypothetical payoffs, which in turn are not different from the baseline real payoff condition. In contrast, Figure 2 demonstrates that for real payoffs subjects switch later to Option B (i.e., they tend to stay with the safer Option A longer), and the higher the stakes, the later they do so; this indicates that people become more risk averse as stakes are increased. In other words, there is an increased tendency to choose the safe option when real stakes are increased.[4]

[4] Harrison, Johnson, McInnes, and Rutstroem (2005) questioned this result arguing that there was an order effect in what Holt and Laury (2002) did and also pointed out that the between-subjects results were problematic because treatments were done at different locations with subject pools that differed on important socio-demographic dimensions. These authors do find their conjecture confirmed, and controlling for socio-demographic characteristics, although they also confirm the key result of Holt and Laury (2002), Harrison et al. (2005) argue that the effect is quantitatively only about half of what Holt and Laury (2002) claimed: "The order effect in the HL design confounds the inference about the scale effects, such that the true scale effect is a little over one-half of the apparent effect when scale and order are confounded. We therefore reaffirm the primary conclusion of HL, that risk aversion varies over the income range found in typical experiments. The effect is significantly smaller than they estimate, but ... does not lead one to reject their qualitative conclusion. Nevertheless, we conclude that order effects are significant and almost as large as scale effects, so that they can lead to misspecification of utility functions." (Harrison et al., 2005, p. 900).

Parco, Rapoport, and Stein (2002). Parco and his colleagues were interested in how the scaling of stakes would affect their subjects' choices in a so-called centipede game. The particular game that they implemented is represented in the following figure, which is reproduced from their article.

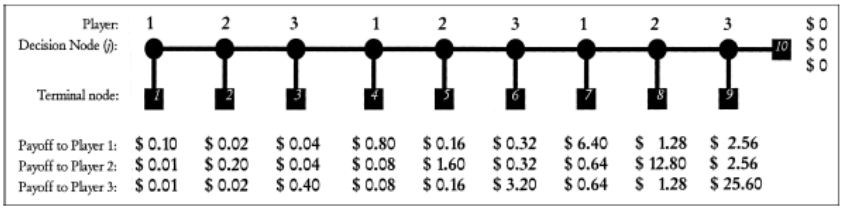

Fig. 3. The three-person, nine-move centipede game used in the present study.

[Fig. 3. reproduced from Parco et al. 2002, p. 293]

In this particular centipede game three players take turns in deciding whether to continue the game or not. The game ends when a player moves down rather than forward. When the game ends, each of the players receives a specific amount of money (shown as "Payoff to Player … " in Fig. 3) which is dependent on the round in which the game was ended. All players know the move structure and the payoffs associated with each decision. The relevant game-theoretic solution concept – subgame perfection – suggests that at the ultimate decision node, a payoff-maximizing player 3 would rather move down than advance since his payoff would be $25.60 rather than $0. Therefore, at the penultimate decision node, player 2 – anticipating player 3's move – would pre-empt that option by herself moving down rather than advancing. By the same reasoning (called backward induction), player 1 would try to pre-empt the later moves of players 2 and 3, and so on. One – somewhat counterintuitive – game-theoretic prediction for this centipede game is therefore that player 1 would move down at the first node, pre-empting other players from making choices. Thus, players would leave considerable amounts of money on the table. Not surprisingly, there have been numerous papers that questioned the logic of backward induction (e.g., Reny, 1992) or that tested the backward induction prediction experimentally (e.g., prominently, McKelvey & Palfrey, 1992). Experimental tests typically found that subjects moved considerably away from the prediction of subgame perfection. That is indeed what Parco and his colleagues found in their experiment.

However, earlier they had conducted the same centipede game with payoffs that were scaled up by a factor of 100 (yes!) and found a dramatic shift towards the game-theoretic prediction. The outcomes of these two treatments are summarized in the table below.

Table 1. *Proportion of games ending at each terminal node*

Session	N^a	Terminal node 1	2	3	4	5	6	7	8	9
Low-pay 9-move game (present study)										
1	300	.027	.043	.093	.240	.263	.227	.073	.017	.013[b]
2	300	.023	.067	.250	.243	.263	.097	.037	.007	.013
Across sessions	600	.025	.055	.172	.242	.263	.162	.055	.012	.013
High-pay 9-move game (RSPN)[c]										
1	300	.463	.317	.110	.050	.027	.020	.010	.003	.000
2	300	.393	.277	.157	.087	.030	.017	.023	.013	.003
3	300	.303	.280	.187	.093	.053	.037	.010	.003	.033
4	300	.407	.257	.183	.077	.037	.017	.013	.003	.007
Across sessions	1,200	.392	.283	.159	.077	.037	.023	.014	.006	.010

[a]Number of games (five groups of 3 randomly matched players per trial participating in 60 trials).
[b]A single Player 3 continued at the 9th decision node.
[c]RSPN = Rapoport, Stein, Parco, and Nicholas (2000).

[Table 1 reproduced from Parco et al. (2002), p. 295]

The relevant comparison is the two lines labeled "Across sessions". For the low-pay condition the terminal nodes with the highest frequency chosen are those for rounds 4 and 5; for the high-pay condition the terminal nodes with the highest frequency chosen are those for rounds 1 and 2. In fact, roughly two thirds of the outcomes are located there and hence rather close to the game-theoretic prediction.

Parco et al. (2002, pp. 295–296) conclude, "The direct comparison of the present low-pay centipede game with the high-pay centipede game ... exhibits strong evidence, perhaps the strongest evidence documented in the literature, that the magnitude of financial incentives makes a significant and substantial difference. Not only do financial incentives matter, but when they are sufficiently high they support Hertwig and Ortmann's (2001) conclusion that when learning is possible, monetary payments may bring the decisions closer to the predictions of the normative models."

Rydval and Ortmann (2004). These authors re-analyzed well-known data from Gneezy and Rustichini (2000), who argued that you ought to pay your subjects enough, or not at all. Specifically, an experimenter may be better off to not pay at all rather than pay amounts that are symbolic at best. We looked more carefully at the disaggregated data of the four treatments – no pay, very low pay, standard pay, high pay – that Gneezy and Rustichini (2000) had undertaken. Under each of these payment conditions Gneezy and Rustichini (2000) had 40 participants take an IQ test, for a total of 160 participants. We ordered the IQ scores for each of the four treatments from lowest to highest, and constructed a "performance" curve for each of these treatments; the four performance curves are shown in the following graph, which is reproduced from the original article.

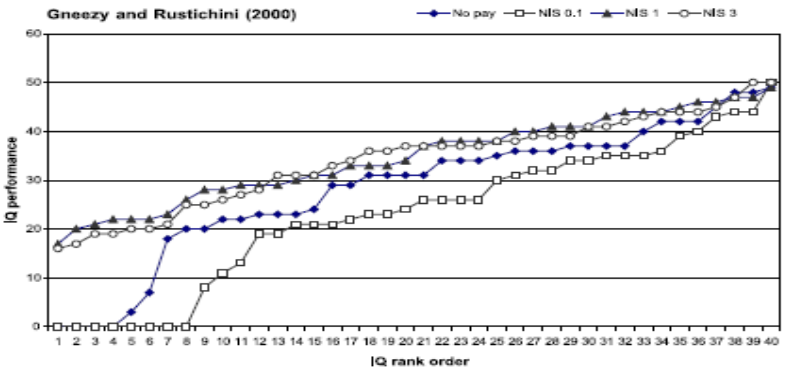

Fig. 1. Individual IQ performance, plotted in ascending IQ rank order, separately for each of the four incentive treatments.

[Fig. 1 reproduced from Rydval and Ortmann (2004), p. 317]

The key observations were: The performance curves for NIS 1 and NIS 3 (the standard- and high-incentive treatments) are almost identical and continuously sloping upward. The performance curves for the very low- (NIS 0.1) and no-pay (the low-incentive treatments) lie below the standard- and high-incentive treatments, the latter less so than the former. This fact contradicts what has been called the labor theory of cognition, which holds that higher financial incentives extract a higher effort in a monotonically increasing manner. (This key result is also what made the Gneezy and Rustichini article a frequently-cited one.)

Importantly, the lower end of the performance curves for the low-incentive treatments reflect either motivational problems or outright sabotage on the part of a non-negligible part of the participants. How can we identify sabotage? Well, to get all of 50 multiple choice questions incorrect, you actually have to exert quite an effort (and be reasonably smart, too!), or simply refuse to answer any question.

Leaving aside these "motivational" problems, it is also noteworthy that within-treatment variation in performance is generally much greater than the variation across treatments (e.g. take the performance differential at the median rank across treatments and compare the within-treatment variation of the 11th-ranked and 30th-ranked). All in all, cognitive abilities seem at least twice as important as financial incentives. This suggests strongly that the traditional emphasis on financial incentives may be very problematic indeed (and that one should in many cases control for cognitive ability).

Experimental results as a consequence of ambiguous language

Moving away from the impact of financial incentives, let us address the importance of saying clearly what the experimental task is. A couple of years ago, Gneezy, List, and Wu published an article (Gneezy et al., 2006) that has since received considerable attention; it seems to question the predictive power of major decision theories such as Expected Utility (EU) Theory or Prospect Theory.

Arguably the most prominent task (of many featured in the article) was the following:

> "Imagine that we offer you a lottery ticket that gives you a 50 percent chance at a $50 gift certificate for Barnes and Noble, and a 50 percent chance at a $100 gift certificate for Barnes and Noble.
>
> Whichever gift certificate you win is good for use within the next two weeks. What is the highest amount of money you would be willing to pay for this lottery ticket?" (Gneezy et al., 2006, p. 1286 and pp. 1301-1302)

In a between-subject design (with hypothetical elicitations of the willingness to pay for the prospects as well as their probabilistic combination by different groups of about 30 subjects), Gneezy and his collaborators find that the hypothetical willingness to pay for this lottery (which involves two gift certificates worth $50 and $100) is actually lower than the willingness to pay for the gift certificate with the lower outcome (i.e., the $50 outcome). This result violates the so-called *Internality Axiom (IA)*, which holds that any probabilistic combination of two such prospects of different values will be valued higher than the lowest prospect (and less than the highest) and indeed in a monotonically increasing manner. That, however, is not what Gneezy and his collaborators find. Here is what they find:

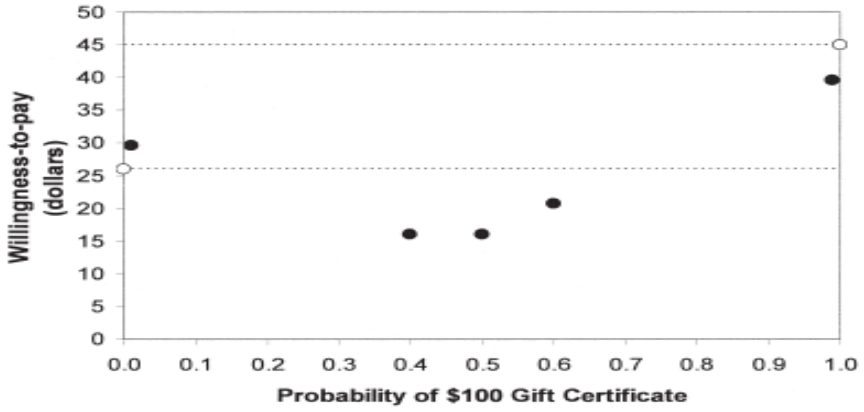

FIGURE II
Willingness-to-Pay (in Dollars) for Gift Certificate Lotteries
Mean willingness-to-pay (in dollars) for various hypothetical lotteries that offer a p chance at a $100 Barnes & Noble gift certificate and a $1 - p$ chance at a $50 Barnes & Noble gift certificate. The open circles indicate sure things or degenerate lotteries, a $50 gift certificate for sure (0 percent) and a $100 gift certificate for sure (100 percent). The filled circles indicate nondegenerate lotteries.

[Figure II reproduced from Gneezy et al., 2006, p. 1288]

One can argue whether experimental between-subjects tests can possibly be damaging for decision theories: It turns out that this result can typically not be produced in within-subject designs, as also acknowledged by the authors. Let us, for the sake of argument, assume that it does make sense to test decision theories with between-subject designs. Clearly, then, the reported *IA* violations would indeed be serious grounds for concern by proponents of EU Theory, Prospect Theory, and many others.

The far-reaching claims by Gneezy and his colleagues deserve a closer look at the way in which the experiment was conducted, which we did in Ortmann, Prokosheva, Rydval, and Hertwig (2007) and Rydval, Ortmann, Prokosheva, and Hertwig (2009). Our initial conjecture was that (some) subjects might not have understood that there

were indeed only two prospects involved. They might, for example, have thought that they were facing a compound lottery involving two lotteries each with two prospects with one lottery involving a *50 percent chance at a $50 gift certificate*, and the other involving a *50 percent chance at a $100 gift certificate*, but both lotteries also having a 50 percent chance of not winning a gift certificate. If indeed that was the case, then subjects might have thought that a zero outcome was possible. (In fact the valuations for the 40/60 and 60/40 mixes seem to be indicative of that.)

We report the test of this conjecture in Ortmann et al. (2007) by first replicating the basic Gneezy et al. treatment for the even 50/50 probabilistic mix ("Replication"). We then reworded the task so as to reduce the possibility of task ambiguity ("Rewording"). Here is what we found:

Table 1: Willingness-to-pay (WTP) in *Replication* and *Rewording*

	Session	WTP (in CZK) in ascending order for each treatment (WTP ≥ 500 in bold)
Replication	1	100 300 300 300 400 400 450 **500 500 500 500 500 500 500 550 600**
	3	250 250 250 300 300 300 300 400 **500 600 600 600 650 730 750 770**
Rewording	2	300 300 375 **500 500 500 500 500 600 600 650 650 700 700 750 750**
	4	250 499 **500 500 500 500 500 500** 550 600 600 620 650 700 700 720

	Replication	*Rewording*
Mean WTP	451.56	555.13
Median WTP	500.00	525.00
Standard deviation of WTP	164.44	129.35
95% C.I. for the means without demographic controls	**(392.28, 510.85)**	**(508.49, 601.76)**
95% C.I. for the means with significant demographic controls	(388.33, 507.28)	(518.41, 599.36)
95% binomial exact C.I. for the medians	(300.00, 500.00)	(500.00, 650.00)
Wilcoxon rank-sum test	Z = 2.56, *p*-value = 0.011	
Kolmogorov-Smirnov test	KS = 0.34, *p*-value = 0.045	
t-test without demographic controls	*t* = 2.80, *p*-value = 0.0068	
t-test with significant demographic controls	*t* = 3.07, *p*-value = 0.0032	
t-test with all demographic controls	*t* = 3.13, *p*-value = 0.0027	

Notes: Raw WTP data (top) are followed by summary statistics and two-sided tests (bottom). Wherever applicable, confidence intervals (C.I.) and tests are based on heteroskedasticity-robust standard errors. The second C.I. pair and the last two *t*-tests are adjusted for the influence of demographics. Session effects, higher-order moments and interactions of demographics, as well as their interactions with the treatments, are individually and jointly insignificant at the 5% significance level.

[Table 1 reproduced from Ortmann et al., 2007]

We found that it seemed to be possible to undo the *Internality Axiom* violation through a rewording of the Gneezy et al. (2006) lottery instructions, at least for the arguably most prominent task in their article. Our argument was based on the observation that the 95% confidence interval was located completely above the face value of the worst possible outcome and that no rational participants would be willing to pay more than the face value of such a gift certificate, which seemed an innocent assumption to make.

Our test illustrates the *semantic* and *pragmatic* ambiguities of natural language, which have been demonstrated amply in the psychology literature (e.g., Hilton, 1995).

In a related vein, Keren and Willemsen (2008) have found that participants who miscomprehend the instructions exhibit an uncertainty effect while participants who do not, do not. Simonsohn (2009) has contested their findings with his own set of experiments (but remains curiously coy on our initial finding, which he cites).

In Rydval et al. (2009), prompted by an invitation to revise Ortmann et al. (2007), we followed up on our earlier study by reducing the possibility of the miscomprehension of the task by physically implementing the lotteries in various ways. Adding many new data to our original pricing task study, and replicating three pricing tasks from Gneezy et al. (2006), which they implemented verbally and for which they found violations of the *Internality Axiom*, we systematically observe that subjects' willingness to pay for the lottery is significantly higher than other subjects' willingness to pay for the lottery's worse outcome.

Do these differences matter?

> "Control is the hallmark of good experimental practice, whether it be undertaken by economists or psychologists." (Harrison & Rutstroem, 2001, p. 413)
> "Finally, the ultimate goal of the laboratory honing of simple models is to explain behavior in the economy." (Camerer, Ho, & Chong, 2004, p. 802)

The answer to each of the following questions is a function of the way one does, and evaluates, an experiment:

- Are people loss-averse? Do people make choices according to PT or other non-EU theories? Do endowment effects exist?
- Are people time-inconsistent? Do people discount options hyperbolically? Do people have a bias toward the present?
- Are people altruistic? Are people fair? Are people reciprocal? Do people trust?

For example, how one evaluates subgame perfection/backward induction (a key theoretical building block tested in experiments on altruism, fairness, reciprocity, and trust), is obviously a function of the stakes involved, or other aspects of design and implementation (see Engelmann & Ortmann, 2009.) Or, as recently suggested by Palacios-Huerta and Volij (2009), it may be a function of the participants' ability to think strategically and their beliefs about the other participants' ability to do so. Likewise, whether one believes that EU Theory and/or Prospect Theory have explanatory value is obviously a function of how serious one takes the results reported in Gneezy et al. (2006).

As it turns out, there are very few experimental results that are not sensitive to the way an experiment is conducted. This implies that experimental control is easily lost when questionable design and implementation choices are made. This poses interesting questions about the appropriateness of experimental design and

implementation choices and the transferability of experimental results to the economy. But this is not the place to address this topic. Unfortunately, a theory of external validity does not yet exist and the problem of external validity remains an interesting and exciting area of dispute and research.

References

Camerer, C. (2003). *Behavioral game theory: experiments in strategic interaction.* Princeton, NJ: Russell Sage Foundation and Princeton University.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: a review and capital–labor–production framework. *Journal of Risk and Uncertainty, 19*, 7-42.

Camerer, C., Ho, T.-H., & Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics, 119*, 861-898.

Engelmann, D., & Ortmann, A. (2009). The Robustness of Gift Exchange. Unpublished Manuscript.

Gneezy, U., List, J., & Wu, G. (2006), The uncertainty effect: When a risky prospect is valued less than its worst outcome. *Quarterly Journal of Economics, 121*, 1283-1309.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics, 115*, 791–811.

Harrison, G. W., Johnson, E., McInnes, M., & Rutstroem, E. (2005). Risk Aversion and Incentive Effects: Comment. *American Economic Review, 95*, 897-901.

Harrison. G.W., & List, J.A. (2004). Field Experiments. *Journal of Economic Literature, 42*, 1009-1055.

Harrison, G.W., &,Rutstroem, E. (2001). Doing it both ways – experimental practice and heuristic context. *Behavioral and Brain Sciences, 24*, 413–414.

Hertwig, R., & Ortmann, O. (2001). Experimental practices in economics: a methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*, 383–451.

Hertwig, R., & Ortmann, O. (2003). Economists' and Psychologists' Experimental Practices: How They Differ, Why They Differ, And How They Could Converge. In I. Brocas and J. Carrillo (Eds.), *The Psychology of Economic Decisions* (pp. 253-272). Oxford: Oxford University Press.

Hertwig, R., & Ortmann, O. (2008a). Deception in Social Psychological Experiments: Two Misconceptions and a Research Agenda. *Social Psychology Quarterly, 71*, 222-227.

Hertwig, R., & Ortmann, O. (2008b). Deception in Experiments: Revisiting the Arguments in Its Defense. *Ethics and Behavior, 18*, 59-92.

Hilton, D.J. (1995). The *social context* of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118*, 248-271.

Holt, C.A., & Laury, S.K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*, 1644–1655.

Keren, G., & Willemsen, M.C. (2008). Decision anomalies, experimenter assumptions, and participants' comprehension: Re-evaluating the uncertainty effect. *Journal of Behavioral Decision Making, 22*, 301–317.

Koehler, J.J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1-53.

Levitt, S.D., & List, J.A. (2007). What do Laboratory Experiments Measuring Social Preferences Reveal About the Real World. *Journal of Economic Perspectives, 21*, 153-174.

List, J.A. (2006). The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy, 114*, 1-37.

McKelvey, R.D., & Palfrey, T.R. (1992). An experimental study of the centipede game, *Econometrica, 60*, 803–836.

Ortmann, A., & Gigerenzer, G. (1997). Reasoning in Economics and Psychology: Why Social Context Matters, *Journal of Institutional and Theoretical Economics, 153*, 700-710.

Ortmann, A., & Hertwig, R. (2002). The Costs of Deception: Evidence from Psychology. *Experimental Economics, 5*, 111-131.

Ortmann, A., Prokosheva, A., Rydval, O., & Hertwig, R. (2007). Valuing a risky prospect less than its worst outcome: Uncertainty effect or task ambiguity? Jena Economic Research Paper 2007-038 and CERGE-EI Working Paper 334.

Palacios-Huerta, I., & Volij, O. (2009). Field Centipedes. *American Economic Review, 99*, 1619–1635.

Parco, J.E., Rapoport, A., & Stein, W.E. (2002). Effects of financial incentives on the breakdown of mutual trust. *Psychological Science, 13*, 292-297.

Reny, P.J. (1992). Rationality in Extensive-Form Games. *Journal of Economic Perspectives,* 6, 103-118.

Rabin, M. (1998). Psychology and Economics. *Journal of Economic Literature, 36,* 11-46.

Rydval, O., Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: an illustration, *Economics Letters, 85*, 315-320.

Rydval, O., Ortmann, A., Prokosheva, A., Hertwig, R. (2009). How certain is the uncertainty effect? *Experimental Economics, 12*, 473–487

Simonsohn, U. (2009). Direct risk aversion: Evidence from risky prospects valued below their worst outcome. *Psychological Science, 20*, 686–692.

Smith, V.L. (1976). Experimental economics: induced value theory. *American Economic Review (Proceedings), 66*, 247–279.

Smith, V.L. (1982). Microeconomic systems as an experimental science. *American Economic Review, 72*, 923–955.

Smith, V.L. (1991). Rational Choice: The Contrast Between *Economics* and *Psychology*. *Journal of Political Economy, 99*, 877-897.

Smith, V.L. (2002). Method in Experiment*:* Rhetoric and Reality*. Experimental Economics, 5*, 91-110.

Smith, V.L. (2008). *Discovery*: *A Memoir*. Bloomington, IN: AuthorHouse.

Smith, V.L., & Walker, J. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry, 31*, 245–261.

Tversky, A., & Kahneman, D. (1987). Rational Choice and the Framing of Decisions. In R.M. Hogarth & M.W. Reder (Eds.), *Rational Choice: The Contrast between Economics and Psychology* (pp. 67-94). Chicago: University of Chicago Press.