CBESS Discussion Paper 09-14

The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: A Reassessment

by Andrea Isoni, Graham Loomes and Robert Sugden *

11 November 2009

*Andrea Isoni: Centre for Behavioural and Experimental Social Science, School of Environmental Sciences and School of Economics, University of East Anglia, Norwich (UK), NR4 7TJ (email: a.isoni@uea.ac.uk); Graham Loomes: Department of Economics, University of Warwick, Coventry (UK), CV4 7AL (email:g.loomes@warwick.ac.uk); Robert Sugden: Centre for Behavioural and Experimental Social Science and School of Economics, University of East Anglia, Norwich (UK), NR4 7TJ (email: r.sugden@uea.ac.uk).

Abstract

Plott and Zeiler (2005) report that the willingness-to-pay/willingnessto-accept disparity is absent for mugs in a particular experimental setting, designed to neutralize misconceptions about the procedures used to elicit valuations. This result has received sustained attention in the literature. However, other data from that same study, not published in that paper, exhibit a significant and persistent disparity when the same experimental procedures are applied to lotteries. We report new data confirming both results, thereby suggesting that the presence or absence of a disparity may be a more complex issue than some may have supposed. (JEL C91, D11).

JEL classification codes

C91, D11

University of East Anglia

CBESS

Centre for Behavioural and Experimental Social Science

CBESS
University of East Anglia
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/ssf/cbess

The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: A Reassessment

Andrea Isoni, Graham Loomes and Robert Sugden*

11 November 2009

* Andrea Isoni: Centre for Behavioural and Experimental Social Science, School of Environmental Sciences and School of Economics, University of East Anglia, Norwich (UK), NR4 7TJ (email: a.isoni@uea.ac.uk); Graham Loomes: Department of Economics, University of Warwick, Coventry (UK), CV4 7AL (email: g.loomes@warwick.ac.uk); Robert Sugden: Centre for Behavioural and Experimental Social Science and School of Economics, University of East Anglia, Norwich (UK), NR4 7TJ (email: r.sugden@uea.ac.uk).

This research was carried out as part of the Programme in Environmental Decision Making, organised through the Centre for Social and Economic Research on the Global Environment at the University of East Anglia, and supported by the Economic and Social Research Council of the UK (award nos. M-535-25-5117 and RES-051-27-0146.) We thank Charlie Plott and Kathryn Zeiler for providing the data set from their experiments, and for their help in designing our experimental procedures. We are grateful to Daniel Zizzo and Chris Starmer, an editor and three referees for comments and suggestions.

Abstract

Plott and Zeiler (2005) report that the willingness-to-pay/willingness-to-accept disparity is absent for mugs in a particular experimental setting, designed to neutralize misconceptions about the procedures used to elicit valuations. This result has received sustained attention in the literature. However, other data from that same study, not published in that paper, exhibit a significant and persistent disparity when the same experimental procedures are applied to lotteries. We report new data confirming both results, thereby suggesting that the presence or absence of a disparity may be a more complex issue than some may have supposed. (JEL C91, D11).

The finding that willingness-to-accept (WTA) measures of value greatly exceed the corresponding willingness-to-pay (WTP) measures has received considerable attention in the last four decades. A large body of evidence, gathered through numerous contingent valuation and experimental studies, shows WTP–WTA gaps of magnitudes that are hard to reconcile with the predictions of standard consumer theory. Surveying forty-five such studies, John K. Horowitz and Kenneth E. McConnell (2002) find that the median ratio of average WTA and average WTP is 2.6 (mean 7.17), as opposed to the few percentage point differences predicted by standard consumer theory (Robert D. Willig, 1976).

These findings have received sustained attention because they seem to conflict with one of the most elementary propositions of consumer theory – that an individual's indifference curves can be specified independently of her endowment and budget constraint. Some theorists have explained the gap as a consequence of a systematic asymmetry between individuals' attitudes to gains and losses relative to some reference point. If such an asymmetry exists, many familiar theoretical results no longer hold. For example, the Kaldor–Hicks compensation test can fail to show direction-neutrality, even if income effects are negligible; and contrary to the Coase theorem, the final outcome of a negotiation may be affected by the initial allocation of property rights, even in the absence of transaction costs. Theories of reference-dependent preferences have been proposed which predict a range of observed deviations from standard consumer theory, including WTP–WTA disparities (Amos Tversky and Daniel Kahneman, 1991; Robert Sugden, 2003; Botond Kőszegi and Matthew Rabin, 2006; Graham Loomes, Shepley Orr and Sugden, 2009; Andrea Isoni, 2009). However, other theoretical explanations have also been proposed, involving, for example, substitution effects (W. Michael Hanemann, 1991), costly information acquisition (Charles D. Kolstad and Rolando M. Guzman, 1999), incompleteness of preferences (Michael Mandler, 2004), and evolutionary advantages (Steffen Huck, Georg Kirchsteiger and Jörg Oechssler, 2005).

In a recent article, Plott and Zeiler (2005) – henceforth PZ – offer a radically different interpretation of disparities between WTP and WTA. The "primary conclusion" they derive from the data they report is that "observed WTP–WTA gaps do not reflect a fundamental feature of human preferences"; the thesis of their paper is that, to the contrary, "observed gaps are symptomatic of subjects' misconceptions about the nature of the experimental task" in which valuations are elicited (p. 542).

PZ's experimental design is grounded in a review of previous experimental investigations of WTP and WTA for a wide range of goods, including low-value consumption goods (such as coffee mugs and chocolate bars), non marketed goods (such as tree density and food safety), lotteries with goods as prizes, and lotteries with money prizes.[1] PZ note that WTP–WTA gaps have been observed in some of these experiments but not in others, and that there is no consensus about the reasons for this variability. They point out that "scholars who accept the psychological explanation of the gap have sought to explain the variation in terms of the commodity used in experiments (e.g., mugs, lotteries, money, candy, etc.)". However, PZ also note that different experiments have used different procedures to try to reduce misconceptions, and they wonder whether these differences in procedures, rather than differences in the experimental commodity, might be the explanation. They argue that, although the literature reveals no agreement about the interpretation of the gap, there is a consensus that experimental procedures to investigate it "should be designed to minimize or avoid subject misconceptions". Because no comprehensive theory of misconceptions exists, they pose the following "main research question": "If we design an experiment that completely controls for subject misconceptions as implicitly defined by the literature (i.e. an experiment that includes every procedure used in previous experiments to control for misconceptions), will we observe a WTP–WTA gap?" (pp. 531–2).

In order to answer this question, PZ run three new experimental treatments in which "subject misconceptions are completely controlled by incorporating the union of procedures found in the literature" (p. 532). They contrast these treatments with a replication of one of a series of experiments reported by Kahneman, Jack L. Knetsch and Richard Thaler (1990) –

[1] PZ's paper gives a brief account of this review. More details are provided in their online Appendix (http://www.e-aer.org/data/june05_app_plott.pdf)

henceforth KKT – which found a WTP–WTA disparity. Summarizing their findings, PZ report that when using the procedures in KKT's Experiment 5, they "replicate the gap with roughly the same magnitude", but their new treatments produce the "striking result" that "[w]hen an incentive-compatible mechanism is used to elicit valuations, and subjects are provided with (a) a detailed explanation of the mechanism and how to arrive at valuations; (b) paid practice using the mechanism; and (c) anonymity," they "observe no WTP–WTA gap" (pp. 531–2). We will use the term *PZ procedure* for the mechanism used in PZ's new treatments. The result stated in the preceding quotation – that no gap is observed when this procedure is used – will be called the *no-gap result*.

Like any experimental finding, the no-gap result has been *demonstrated* only in the confines of specific experimental designs, but is *significant* by virtue of interpretative judgements about its wider applicability. Describing the domain of their own experiments, PZ state: "Experiments were conducted using both lotteries and mugs, goods frequently used in endowment effect experiments. Using the modified procedures, we observe no gap between WTA and WTP" (Abstract, p. 530). Interpreting their result, they conclude: "The differences reported in the literature reflect differences in experimental controls for misconceptions as opposed to differences in the nature of the commodity (e.g., candy, money, mugs, lotteries, etc.) under study" (p. 542; see also p. 531).

Possibly because of the reference in the Abstract to the use of two goods, and because PZ suggest that the variation in experimental results is related to variation in controls for misconceptions rather than variation in the goods being used, their paper has been widely cited as providing experimental support for the hypothesis that the PZ elicitation procedure eliminates WTP–WTA disparities *in general* and that such disparities are artifacts of insufficiently

controlled experiments. For example, Jay R. Corrigan and Matthew C. Rousu (2008, p. 291) interpret PZ as showing "that the widely heralded disparity between [WTA] and WTP may simply be an artifact of participants' 'misconceptions' regarding the demand-revealing nature of widely used auction mechanisms". Matan Tsur (2008, p. 740) takes the PZ paper as evidence "that subjects' misconceptions of experimental tasks are the main cause for the WTA/WTP disparity reported in experiments". Many similar readings of PZ can be found in recent literature.[2] In a later paper, PZ themselves refer to their study as follows: "Plott and Zeiler (2005) posited an explanation centered on subject misconceptions stemming from the preference elicitation method, and ran additional experiments that implemented the union of commonly used controls to reduce misconceptions. When procedures were used to eliminate alternative explanations, the gap disappeared. The data support the conclusion that observed WTA–WTP gaps are caused by subject misconceptions resulting from the use of special mechanisms required to elicit valuations." (Plott and Zeiler, 2007, p. 1450).

Given the amount of effort that has been and continues to be devoted to theoretical explanations of WTP–WTA disparities, the hypothesis that these disparities are mere artifacts is controversial and potentially of great significance. However, PZ's published results provide a somewhat limited evidential base for this hypothesis: although their experiments involved 14 tasks eliciting WTP and WTA for lotteries and one eliciting WTP or WTA for a mug, their paper reports only the results from the task involving mugs. Thus their no-gap result rests on 36 subjects reporting WTP for a coffee mug and 38 reporting WTA.

---

[2] See, for example, Jeffrey J. Rachlinski (2006), Stelle N. Garnett (2006), Eric J. Johnson, Gerald Häubl and Anat Keinan (2007), Emmanuel Flachaire and Guillaume Hollard (2008), and Jeremy Clark and Lara Friesen (2008, p. 197).

PZ explain the exclusion of their lottery data on the grounds that these tasks were used only for training subjects and that this may have led to data contamination (pp. 539–40, note 15). However, given the widespread use of lotteries in economics experiments, it is of interest to ask whether, in the absence of contamination, PZ's elicitation procedure eliminates the WTP–WTA gap for lotteries as well. If the answer is 'Yes', that would be consistent with the now-common interpretation of PZ's results as showing that WTP–WTA gaps in general are artifactual; but if the answer is 'No', we conjecture that researchers might be more cautious about dismissing the WTP–WTA disparity simply as an artifact or supposing that it can be *generally* eliminated by those particular procedures.

In mounting our study, our primary objective was to apply PZ's elicitation procedure to both mugs and lotteries while ensuring that none of the paid tasks was contaminated. As in PZ's experiment, we found no significant WTP–WTA gap for mugs, thereby adding weight to that particular result. However, we also observed a significant and persistent gap for lotteries of much the same kind found in PZ's unreported lottery data, suggesting that the PZ procedure does not *in general* eliminate the WTP–WTA gap.[3]

In a second stage of our investigation, we compared the KKT and PZ designs when applied to mugs.[4] By contrast with PZ, who used different subject pools and different types of mugs in their various treatments, we implemented a controlled comparison in which subjects

[3] We acknowledge that PZ have made their full data set available to other researchers on request, and that the notes that accompany these data refer to the WTP–WTA gap for lotteries. But since there is nothing in the published paper or the online Appendix to suggest that these data might reveal this result, that result cannot be said to have been reported in the normal scientific sense.

[4] This comparison was requested by two referees of an earlier version of our paper.

were randomized between the two treatments, with the same mugs being used in both. Moreover, whereas PZ gave a show-up fee in their own treatments but not in their KKT replication, we gave the same show-up fee in both treatments. Under these conditions, we found no significant differences between the two procedures. This finding does not affect our conclusions about the no-gap result, but it raises questions about how far the WTP–WTA disparity for mugs is attributable to subjects' misconceptions and how far other variables in the experimental design might be influential.

Before moving to the substantive part of our paper, we must explain its intended scope. Like PZ, we believe it is important to distinguish between the *existence* of the WTP–WTA gap as an empirical regularity and possible *explanations* for that phenomenon. PZ emphasize this difference by using the term "WTP–WTA gap" to refer to the empirical regularity and by introducing the term "endowment effect theory" (EET) to refer to "a particular explanation of gaps", namely that they are due to loss aversion (pp. 530–32, 542). PZ present their explanation of the WTP–WTA gap – that it is the result of subject misconceptions – in opposition to the hypothesis that it is explained by EET. They conclude that EET "does not seem to explain observed gaps" (p. 542).

PZ do not set out EET as a specific formal theory, saying only that it is "a special theory of the psychology of preferences associated with 'prospect theory'" (p. 531). They do not cite any theoretical presentation of prospect theory, instead using KKT's primarily empirical paper as their main point of reference. As used by PZ, "EET" appears to refer to a loosely-related family of theories of reference-dependent preferences which has evolved and diverged over time.[5] Thus,

[5] For example, prospect theory as originally proposed by Kahneman and Tversky (1979) applies only to pairwise choices under risk in which the reference point is some sure amount, and so makes no predictions about WTA for

whether any particular testable hypothesis is an implication of EET will often be a matter of judgment. Implicitly, PZ attribute two firm hypotheses to EET. First, there is some tendency for WTA to exceed WTP. Second, because this disparity results from a "fundamental feature of human preferences", it cannot be "turned off and on using different sets of procedures". More specifically, it cannot be turned off simply by using procedures which control for subject misconceptions (pp. 531–532). By attributing these hypotheses to EET, PZ are able to interpret their "main research question" as a test of EET and hence to interpret the no-gap result in the case of mugs as evidence against that theory.

If KKT are regarded as the principal exponents of EET, the attribution of the second hypothesis to that theory might be questioned.[6] In this paper, however, we are *not* concerned

---

lotteries, nor about either WTP or WTA for riskless objects such as mugs. Tversky and Kahneman (1991) use the assumptions that underpin the value function in prospect theory and apply these to choices between riskless goods in such a way as to imply WTP–WTA gaps for such goods; but this model is silent about lotteries. By contrast, Sugden's (2003) reference-dependent expected utility theory predicts a WTP–WTA gap for lotteries.

[6] KKT do not hypothesize that WTP–WTA gaps *always* occur and necessarily persist even when other causes are controlled. For example, they refer to situations where "the buying-selling discrepancy is simply a strategic mistake, which experienced traders will learn to avoid" (p. 1326). Their most explicit statement of a theoretically-grounded hypothesis is in the following passage: "[M]any discrepancies between WTA and WTP, far from being a mistake, reflect a genuine effect of reference positions on preferences. Thaler (1980) labeled the increased value of a good to an individual when the good becomes part of the individual's endowment the 'endowment effect'. This effect is a manifestation of 'loss aversion', the generalization that losses are weighted substantially more than objectively commensurate gains in the evaluation of prospects and trades (Kahneman and Tversky, 1979; Tversky and Kahneman, in press [1991]). An implication of this asymmetry is that if a good is evaluated as a loss when it is given up and as a gain when it is acquired, loss aversion will, on average, induce a higher dollar value for owners than for potential buyers, reducing the set of mutually acceptable trades" (pp. 1327–8).

with PZ's interpretation of what they call EET. Nor are we concerned with whether PZ's or our own data are consistent with EET. Since EET is not a sharply-defined concept, engagement with those issues would be an unhelpful distraction from the point of our paper. Indeed, we are not concerned with testing *any* particular theoretical account of WTP–WTA gaps. We focus on two better-defined questions relating to the *existence* of the WTP–WTA gap: first, whether it is *generally* eliminated when the PZ elicitation procedure is implemented; and second, whether the WTP–WTA gap for mugs is less evident when the PZ procedure is used rather than the KKT design and when other potentially confounding variables are better controlled.

The remainder of this paper is organized as follows. In Section I we describe the main features of PZ's design and argue that, if contamination problems are avoided, the lottery tasks are well-suited for exploring the relationship between WTP and WTA. In Section II we describe our uncontaminated replication of PZ's experiment. In Section III we report the results of our experiment and compare these with the corresponding data from PZ's. In Section IV we offer some conjectures about possible causes of the difference we observe between mug and lottery tasks, and about why there is no significant WTP–WTA gap in our replication of the KKT experiment. But, of course, these can only be post hoc speculations, although we hope they may suggest possible topics for future experimental investigation.

I. Plott and Zeiler's design

In this Section, we focus on those parts of PZ's experiment that implement the PZ elicitation procedure via their Treatments 1, 2 and 3.[7] The overall structure of each of these treatments consists of the following sequence of phases: (i) general instructions, (ii) worked examples, (iii) unpaid training rounds, (iv) paid rounds, and (v) payments. The general instructions explain the elicitation mechanism, a variant of the Becker–DeGroot–Marschak (BDM) procedure (Gordon M. Becker, Morris H. DeGroot and Jacob Marschak, 1964). Numerical examples are then used to show why, when this procedure is used, it is optimal to report a WTP or WTA equal to one's true value. In the course of this phase, participants are shown hypothetical WTP and WTA tasks (each involving lotteries, but with outcomes represented by pure numbers rather than amounts of money) and are given instructions about how to enter valuations on the forms used to record responses. In the unpaid training rounds, participants work through two hypothetical tasks (one WTP and one WTA) involving lotteries with money outcomes. Participants are free to ask questions, and mistakes are identified and corrected by the experimenters. WTP is elicited for a degenerate lottery, offering a small sum of money with certainty. The experimenter uses this task to reinforce the message that a participant who fails to report his or her true value (which in this case is unambiguous) is liable to make avoidable losses. The fourth phase contains the 15 paid tasks, in which WTP and WTA valuations are elicited for 14 different lotteries and for a mug.[8] In

[7] PZ refer to these as 'Experiments' 1, 2 and 3. Treatment 3 (carried out several months after the first two treatments and with a different subject pool) included debriefing interviews. Findings from these interviews are discussed in Plott and Zeiler (2002).

[8] In their paper, PZ say very little about the results generated by the lottery tasks, apart from noting some "speculations and conjectures" based on their interpretation of the unreported data (pp. 539–540, note 15). In the

the final phase, participants receive the net payments to which they are entitled as a result of the paid tasks, in addition to a show-up fee of $5.00. The payment procedure is organized so that each participant's payout is not known by other participants or by the experimenters.

In Treatments 1 and 3, the sequence of paid tasks consists of 14 lottery tasks, described by PZ as "paid practices", followed by a mug task. The lottery tasks are sequenced as follows: three tasks elicit WTA for small-stake lotteries, two of which are degenerate; three more tasks elicit WTP for small-stake lotteries, two of which are degenerate; then four tasks elicit WTA for (relatively) large-stake, non degenerate lotteries; after which, four tasks elicit WTP for (relatively) large-stake, non degenerate lotteries. For these lottery tasks, participants are allocated randomly between two groups, A and B, with the parameters of these tasks differing between the two groups. For the mug task, participants are randomly allocated between WTP and WTA. Treatment 2 is identical, except that the mug task precedes the sequence of lottery tasks.[9] In every task, each participant reports his valuation and then observes the realization of the BDM procedure which determines his payment for that task; these payments are accumulated over the course of the experiment and paid out at the end.

The parameters of the PZ lotteries are shown in the final two columns of Table 1. Notice that each of the lotteries for which WTP is elicited is obtained by adding $0.10 (for small-stake lotteries) or $1.00 (for large-stake lotteries) to the corresponding WTA lottery. For example,

---

online Appendix (see note 1 above), they describe the lottery tasks in more detail, but not the data that they generated. We thank them for giving us access to these data.

[9] In the paper and online Appendix, PZ do not explain the purpose of the lottery tasks in Treatment 2. In a private communication, they have informed us that in this treatment the lottery tasks were used as an "exploratory experiment" to gain insights into subjects' misconceptions.

lottery 3 for group A, denoted by ($0.70, 0.3; –$0.20, 0.7), is a low-stake WTA lottery which gives a gain of $0.70 with probability 0.3 and a loss of $0.20 with probability 0.7. The corresponding WTP lottery is lottery 6, i.e. ($0.80, 0.3; –$0.10, 0.7), which is obtained from lottery 3 by adding $0.10 to each outcome. In general, if $L = (x, p; y, 1 - p)$ is a WTA lottery, the corresponding WTP lottery can be written as $K = (x + c, p; y + c, 1 - p)$.

[Table 1 about here]

This feature of the design is particularly well-suited for making within-subject comparisons of WTP and WTA. Under the assumption of constant absolute risk aversion, expected utility theory implies WTP(K) – WTA(L) = c (where WTP(K) and WTA(L) denote WTP and WTA valuations of the respective lotteries). For any credible assumptions about participants' wealth levels and about the curvature of their utility-of-wealth functions, the degree of approximation involved in assuming constant absolute risk aversion is tiny, even allowing for changes in participants' wealth over the course of the experiment as a result of the accumulation of payoffs from successive tasks.[10] Further, any (small) effects associated with this approximation can be expected to work in a consistent direction. It is a standard assumption in expected utility theory that absolute risk aversion falls as wealth increases. Because the increment c is added to WTP lotteries rather than to WTA lotteries, and because WTP valuations are always elicited *after* the corresponding WTA valuations (that is, when participants' accumulated earnings are expected to be higher), any wealth effects will tend to make WTP higher than it would be under the

[10] The theoretical justification for this claim is provided by Rabin's (2000) 'calibration theorem'. The argument assumes that utility is defined on *levels* of wealth rather than on *increments*; but in the present context that is not a problem. The hypothesis that utility is a function of increments of wealth is a hypothesis about the reference-dependence of preferences, while PZ's null hypothesis is that preferences are reference-*independent*.

assumption of constant absolute risk aversion. Thus, that assumption imparts a slight conservative bias to a test for disparities in which the null hypothesis is WTP(K) – WTA(L) = c and the alternative is WTP(K) – WTA(L) < c. In this respect, the non counterbalanced order of WTA and WTP tasks serves a useful purpose.[11]

If one considers the structure of the experiment as we have described it so far, the lottery tasks seem no less well-designed than the mug tasks for testing hypotheses about the existence of WTP–WTA disparities. Indeed, whereas the mug tasks here require between-sample tests, the lottery data allow more powerful within-subject tests. In design terms, there seems to be no prima facie reason to treat the mug task as the "real" test for disparities and the lottery tasks as "paid practices" (when preceding the mug task) or as irrelevant (when coming after), any more than the converse. From the participant's point of view, all paid tasks are "real" in the relevant sense of having real consequences. We suggest that hypotheses about the effects of paid practice can more usefully be formulated as hypotheses about the effects of *experience* – that is, of repeatedly facing paid tasks, none of which needs to be described to subjects as a practice for something else. As PZ say at the beginning of their footnote 15: "Theoretically, the lottery rounds could be used to test for a WTP–WTA gap" (p. 539).

However, PZ did not consider it appropriate to use their lottery data for testing hypotheses, on the grounds that these data were "contaminated by a design that was developed

[11] One problem (pointed out by PZ in a private communication) is that, if misconceptions are eroded gradually, WTA responses will be more affected by misconceptions than WTP responses. However, if such learning were taking place, one would expect WTP–WTA disparities to be greater for pairs of tasks that appear earlier in the experiment than for ones that appear later; and that can be investigated. In fact, there are no obvious trends in the degree of disparities for lotteries, either in PZ's experiment or in our replication of it (see Section III).

only for training". They refer to two forms of "contamination". First, lottery selling tasks preceded lottery buying tasks. Second, "[m]istake corrections, public answers to questions, and other procedures were also employed continuously, which confound the valuations provided in the lottery rounds" (PZ, pp. 539–540, note 15). We have already explained why the order in which buying and selling tasks were presented does not rule out tests for WTP–WTA disparities. We now consider the role of possible contamination of subjects' responses to the lottery tasks by training procedures.

The written instructions for PZ's experiment provide very little information about the content of these training procedures.[12] The only reference to any practice or training associated with the paid lottery tasks comes at the very end of the instructions, where subjects are told: "the first several rounds involve relatively small payoffs. These rounds are intended to give you practice before you get to the rounds involving significant payoffs". This passage seems to be advising subjects to use the small-stake lottery tasks as practices for the large-stake lottery tasks and possibly (in the case of Treatments 1 and 3) also for the mug task, but it gives no indication of the nature of any intervention by the experimenters in any of the lottery tasks. The only additional information in PZ's paper and online Appendix is the passage we have already quoted. From this information, it is difficult to judge how intrusive (and thereby potentially contaminating) the experimenters' interventions were.

However, since any analysis of PZ's lottery data would be vulnerable to the criticism that these data might have been contaminated, we decided to replicate Treatment 1 (and Treatment 3, which is identical except for the ex-post debriefing) but with the crucial difference that none of

[12] These instructions are reproduced in PZ's online Appendix: see note 1 above.

the paid tasks was described as a "practice" and that, when these tasks were being performed, there was no training intervention by the experimenters. In other respects, we tried to replicate PZ's procedures as closely as possible.

II. Our design

In Stage 1 of our study, our aim was to replicate PZ's implementation of their elicitation procedure for lotteries and mugs while ensuring that the lottery data were not contaminated.[13] Most of the differences between the original design and the replication are adaptations necessary for a computerized implementation, rather than PZ's pen-and-paper methods. We chose to use computers to simplify the organization of the experiment[14] and to make the interface between participant and experiment as pre-scripted as possible.

Our experiment had the same five phases as the original. In the instruction phase, the instructions reproduced those of the original experiment very closely, with a slightly different but entirely standard visual representation of lotteries. They were read out by an experimenter while participants followed the text in printed form. The full text of our instructions can be found in the Appendix [intended only for online publication].

---

[13] We thank Kathryn Zeiler for her assistance in the design of the experiment and in the preparation of the experimental instructions.

[14] Computerization avoids the need for PZ's complex and time-consuming "commitment" procedure (in which subjects write their WTP or WTA valuations on slips of paper and post them in sealed boxes, and the experimenters subsequently check these against subjects' record sheets). In the training rounds, computerization also simplifies the checking of responses for inconsistencies.

In the "worked examples" phase, participants were shown two valuation tasks, one eliciting WTP for a non degenerate lottery and one eliciting WTA for a degenerate lottery. For each of these examples, they were shown the five steps that would later be followed in each unpaid practice and in each paid task. In Step 1, they would enter (open-ended) valuations, rounded to the nearest five pence. In Step 2, the experimenter would reveal the fixed offer by publicly opening a colored envelope randomly selected from a set of 80.[15] In Step 3 (for lotteries only), the outcome would be publicly determined by drawing one of 100 numbered discs from a bag. Participants would record the monetary outcome corresponding to the drawn number, which could be read easily from the lottery display on the screen. In Step 4, participants would work out and enter their net earnings for the round; the program would then verify these entries. In Step 5, they would add these earnings to (or subtract them from) the accumulated total of previous rounds; the program would verify the new total.

The training phase involved two unpaid tasks. These were exactly as in the PZ treatments, except that lottery outcomes were expressed in UK pounds instead of US dollars. In the first training round, participants reported their WTP for the degenerate lottery (£3, 0.7; £3, 0.3), while in the second they reported their WTA for (£2, 0.5; £4, 0.5). In the training phase (but not in the later paid tasks), whenever a subject entered a value outside the range of possible payoffs, the computer displayed an error message explaining why the value was not optimal. Before proceeding, the experimenter clarified any doubts regarding the message on the screen

[15] Two sets of colored envelopes were used, one for the first six tasks and the other for the later tasks. All offers were in multiples of five pence. The distribution of offers, different for the two sets, was not revealed to participants.

and answered any questions. Subjects who had entered non optimal values were given a chance to revise their valuations if they wanted to.

There were 16 paid tasks. The first 15 of these were very similar to the 15 tasks of the PZ treatments, with the lottery tasks presented first (as in PZ's Treatments 1 and 3). In the interests of statistical power, we did not distinguish between type A and type B lotteries as PZ did: all subjects valued the same lotteries (and in the same order). The parameters of these lotteries are shown in the 'Replication lottery' column of Table 1. After allowing for a conversion rate (at the time of the experiment) of approximately two dollars to one pound, these parameters are broadly similar to those used by PZ, except that the payoffs in our small-stake lotteries are somewhat larger than in PZ's. Just as in the PZ experiment, each WTP lottery is constructed from a corresponding WTA lottery by adding a constant amount to each outcome (£0.10 for small-stake lotteries, £1.00 for large-stake lotteries). For consistency with the recruitment methods and experimental practices that are standard at our lab, we did not include lotteries involving losses; this required us to create substitutes for PZ's lotteries 3, 6, 9 and 13. For the fifteenth task, participants were divided between WTA and WTP treatments; valuations were elicited for a University of East Anglia coffee mug (with a retail price of £4.50).

The final paid task was new to our experiment. This task elicited valuations of a *chocolate gamble* (CG) offering a 0.25 probability of winning a box of luxury chocolates (with a retail price of £13.50) and a 0.75 probability of winning nothing.[16] Participants who reported

[16] Unlike the other lotteries, which were played out publicly during the experiment (with the same realisation of the random process for all participants in a session), the chocolate lottery was played out separately for each participant who bought or failed to sell. This procedure was used to reduce the experimenters' ex ante uncertainty about how many boxes of chocolates would be required for each session.

WTA in the mug task reported WTP in the CG task, and vice versa. We introduced this task because, in view of the PZ data, we conjectured that the extent of disparities might differ between lottery and mug tasks. Such a pattern might be explained as the effect *either* of a difference between lotteries and certainties *or* of a difference between money outcomes and outcomes described in terms of consumption goods. By eliciting valuations for a gamble with a consumption good as a prize, we hoped to throw some tentative light on this issue. Since participants faced the 15 PZ tasks before even being aware of the existence of the CG task, the latter could not contaminate our replication.

The payment phase was designed to replicate the anonymity of the PZ experiment as far as possible, subject to the constraint that we are required by tax regulations to collect signed receipts from people taking part in our experiments. Anonymity was implemented as follows. An assistant checked participants' identity on arrival at the lab. The experimenter inside the lab was unaware of the names of the participants, each of whom was identified by a unique 7-digit identification code contained in a sealed envelope. At the end of the experiment, participants left the lab and received their earnings (including a £3.00 show-up fee) at a pre-specified time and place from a cashier, who asked them to sign a receipt and withdrew their identification card. As the instructions explained, this ensured that the cashier (who had no other connection with the experiment) was the only person able to associate individual participants with their payoffs.

Stage 2 of our investigations was a controlled comparison between the PZ and KKT designs. From now on, we will call PZ's replication of the KKT design the "KKT–PZ treatment" and our replication of it the "KKT–ILS treatment"; our Stage 2 treatment using the PZ elicitation procedure will be called the "PZ–ILS" treatment. PZ use the KKT–PZ treatment as a benchmark

against which to measure the effectiveness of their controls for misconceptions. In order for this to be an informative comparison, however, the two sets of procedures should be as comparable as possible. For this reason, our PZ–ILS and KKT–ILS treatments differed only with respect to what PZ regard as their essential controls for misconceptions. In order to achieve this, we took the following steps. The participants were not the same as those in Stage 1, but they were recruited from the same subject pool, and were randomly divided between the PZ–ILS and KKT–ILS treatments. Each treatment elicited WTP and WTA for a mug, the same type of mug in both treatments.[17]

Our PZ–ILS treatment was essentially the same as our Stage 1 experiment, except for four modifications. First, there were no lottery tasks in the "paid task" phase; participants moved straight from the training phase to the mug task.[18] In this respect, the status of mug tasks in the PZ–ILS treatment was similar to that in PZ's Treatment 2 (in which the mug task was the first paid task), which we had not replicated in Stage 1. Second, the "worked example" and training phases used tokens with fixed redemption values instead of degenerate and non degenerate lotteries. Third, a mug was placed in front of every participant, as in the original PZ experiment. (In Stage 1, as part of our computerization of the design, we had substituted an on-screen

[17] In these respects, PZ's comparison between the two designs was less controlled. The treatments which used the PZ elicitation procedure used students at the University of Southern California Law School (Treatments 1 and 2) and Pasadena City College (Treatment 3); the KKT–PZ treatment used students at CalTech. The mug used in the KKT–PZ treatment was different from that used in Treatments 1, 2, and 3.

[18] Given that no lotteries were used and that there was only one paid task, in each round subjects had to complete only three steps: entering their offer, recording the fixed offer, and computing their round payment.

photograph of a mug.) Finally, we increased the show-up fee from £3.00 to £8.00 to compensate for the absence of the lottery tasks.

As in the KKT–PZ treatment, the KKT–ILS treatment elicited hypothetical WTP or WTA valuations for two fixed-value tokens (the same tokens as in the PZ–ILS treatment), prior to the mug task.[19] Buyers and sellers sat in adjacent seats; a mug was placed in front of each seller, and buyers could inspect this. In the interests of greater comparability with our PZ–ILS treatment, we made two changes. First, our implementation was computerized. Second, we paid the same £8.00 show-up fee as in the PZ–ILS treatment. (The KKT–PZ treatment, like the original KKT experiment, had no show-up fee.) This latter change was introduced in order to control for a potentially confounding factor. It is possible that subjects' responses to valuation tasks are affected by their previous experimental earnings, in the form of show-up fees and earnings from previous tasks. In particular, we could not be sure that responses would be immune from "house money" effects (Richard Thaler and Eric Johnson, 1990) that might exert an upward pressure on WTP responses while leaving WTA relatively unaffected. Since any such effects would attenuate WTP–WTA disparities, we considered it desirable to control for this possibility when making comparisons between the PZ procedure and the KKT design.

III. Results

Both stages of the experimental investigation were conducted at the Social Science for the Environment Experimental Laboratory of the Univerity of East Anglia using the Zurich Toolbox

[19] The full instructions of the KKT–ILS treatment are reported in the Appendix.

for Readymade Economic Experiments (Urs Fischbacher, 2007). In total we recruited 244 subjects – 100 for Stage 1 and 144 for Stage 2 – drawn from the general student population.

The results are presented in Table 2 below, which also reports PZ's data for comparison. Each column in the table refers to a matched pair of WTA and WTP tasks, either within-subject (for lotteries) or between-subject (for the mug and CG). In the column headings, L1 to L14 refer to lotteries 1 to 14 in the relevant experiment or treatment. For each of the two tasks, the table shows: the number of observations (*n*); (for lottery tasks) the expected value (EV) of the lottery; the mean, median and standard deviation of participants' reported valuations; and (for lottery tasks) the ratio of the mean reported valuation to the EV. For each pair of tasks, the table shows mean and median 'standardized WTA/WTP' statistics. For the lottery tasks, standardized ratios are defined as [WTA(L) + c] / WTP(K);[20] the statistics reported are the means and medians of the within-subject ratios. For the between-subject mug and CG tasks, we report the ratio of mean WTA to mean WTP and the ratio of median WTA to median WTP. The final row reports the result of a test of the hypothesis that, after standardization, WTA is greater than WTP.[21] For

[20] Relative to the obvious alternative, namely WTA(L) / [WTP(K) – c], this definition gives lower values and is compatible with observations for which WTP(K) ≤ c.

[21] When offers are constrained to take non negative values, a truncation problem may arise every time the minimum prize is zero or less (as in WTA lotteries 3, 7, 8, and 10 of the replication experiment and also lotteries 6, 9 and 13 of PZ's experiment). The essence of the problem is that errors that would make WTA lower than zero are ruled out in these cases, potentially creating artificial WTP–WTA disparities. However, if truncation were a serious issue, one would expect a large number of zero valuations for these lotteries. Since this is never the case in our data, and occurs extremely rarely in PZ's data, we can be confident that our tests are capturing genuine WTP–WTA disparities.

lottery tasks, the significance level reported in the last row is for a one-tail Wilcoxon signed-rank test, while for other tasks it is for a one-tail Mann-Whitney test.

[Table 2 about here]

Before considering the main results, we look at the degenerate lottery tasks (rounds 1, 2, 4 and 5 of Stage 1). Given that participants' and fixed offers were constrained to be multiples of five pence, each of these tasks had two responses consistent with a weakly dominant bidding strategy, namely x and $x + 0.05$ in WTA tasks and x and $x - 0.05$ in WTP tasks (where x is the certain amount). Averaging over the four tasks, 77.3 percent of responses satisfied this criterion, and 86.5 percent were within five pence of this; there was no particular trend. 60 percent of subjects made weakly dominant bids in all four tasks, while only 6 percent made dominated bids throughout. The frequency of dominated bids was higher than in the original PZ experiment, but the two are not comparable: we did not deploy any forms of mistake correction at this stage.

We now turn to the non degenerate lottery tasks (i.e. tasks 3 and 6–14). In our experiment, as shown in the last row of panel A of Table 2, WTA significantly exceeds WTP at the 1 percent level in four of the five possible comparisons.[22] Panels C and D show a very similar pattern in the PZ experiments, where WTA significantly exceeds WTP in all ten comparisons (at the 1 percent level in six cases and at the 5 percent level in the others). In both sets of data, standardized WTA/WTP ratios are somewhat lower than in most comparable studies (ranging from 1.11 to 2.19 in our experiment and from 1.13 to 1.97 in PZ's), but the existence of

[22] The only case in which the disparity is not statistically significant is the pair of lotteries 9 and 13. It may be relevant that this is the only case, either in our experiment or in Plott and Zeiler's, in which the selling lottery is non degenerate and has two positive outcomes.

the disparity is absolutely clear. The strong similarity between the two sets of results suggests that the training procedures that accompanied PZ's lottery tasks did not induce systematic distortion.

Is there any tendency for the extent of the disparity to decay as participants gain experience? Since the WTA tasks were presented in the same order as the corresponding WTP tasks, we can investigate this question by looking for trends in the standardized WTA/WTP ratios over the sequence of lottery pairs (3, 6), (7, 11), (8, 12), (9, 13) and (10, 14). In each of the three data sets there some variability, but looking at the data as a whole, this variability appears to be essentially random. WTA valuations (which are reported around the middle of each experimental session) show a consistent tendency to exceed EVs (the ratio of mean WTA to EV is greater than 1 in 11 cases out of 15), while WTP valuations (mostly reported towards the end of the session) show a similarly consistent tendency to fall short of EVs (the ratio of mean WTP to EV is less than 1 in 11 cases out of 15).

Finally, we consider the mug tasks. Panel E of Table 2 shows the data reported by PZ in support of their no-gap result. The key finding is that WTA is not significantly greater than WTP. (In fact, and quite unusually, WTA is *less* than WTP.) This is the case both when the mug task comes after the lottery tasks (Treatments 1 and 3) and when it comes before (Treatment 2). The results of our replication are shown in panel A of Table 2. We find a small positive disparity – the ratio of mean WTA to mean WTP is 1.19 – but this is not statistically significant. Again, there is an obvious similarity between the results of the original experiment and of the replication. The absence of any disparity for mugs when the PZ procedure is used is also evident in the results of Stage 2, which are reported in panel B of Table 2. There is no significant

difference between the distributions of WTA and WTP valuations; the ratio of mean WTA to mean WTP is 0.90 (1.20 for medians).

We find similar results when the KKT procedures are used. Here too there is no significant difference between the distributions of WTA and WTP; the ratio of means is 0.96 (1.22 for medians). Recall that, in Stage 2, participants were randomized between the PZ–ILS and KKT–ILS treatments, the same mug was traded in each treatment, and the show-up fee was the same. Thus, our data (unlike PZ's) permit controlled comparisons of valuations across treatments. We find no significant cross-treatment differences, either for WTA or for WTP (see note c in Table 2).

IV. Discussion

Our primary conclusion is that PZ's no-gap result does not hold for (monetary) lotteries, but does hold for mugs. In PZ's treatments, and in our Stage 1 replication, the procedures for eliciting valuations are essentially the same for both lotteries and mugs. If WTP–WTA disparities were produced simply by misunderstandings of elicitation procedures, and if the variation in the extent of these disparities found in the literature were attributable to differences in controls for misconceptions, we would expect the elimination of disparities in valuations of one good to be associated with the elimination of disparities in valuations of others. It is not credible to propose that misconceptions about a common set of elicitation procedures persist, without any obvious tendency for decay, over a series of paid lottery tasks, and then suddenly disappear when the mug task is faced. And this kind of explanation clearly cannot rationalize the pattern found in PZ's Treatment 2, where the disparity is absent in the first paid (mug) task and then appears and persists over a sequence of later (lottery) tasks.

If one looks only at PZ's own data, the existence and persistence of the WTP–WTA disparity for lotteries is clearly a systematic effect. Since one might expect that mistake correction procedures would, if anything, tend to *reduce* the effect of misconceptions, it is hard to see how the persistent disparity in the PZ lottery data could be an artifact of contamination from this source. And our replication shows that the disparity continues to be observed when that source of potential contamination is removed. The obvious inference to draw is that, when the PZ elicitation procedure is used, the WTP–WTA disparity *is absent for mugs but occurs and persists for lotteries*.

As we explained in the introduction, although PZ's no-gap results come only from the one round in each of their treatments involving mugs,[23] their contribution has been widely misinterpreted as demonstrating that WTP–WTA gaps more generally are artifacts of elicitation procedures that fail to correct respondents' misconceptions. The fact that mugs are a staple commodity in WTP–WTA experiments, the wording in the PZ abstract and the absence of conflicting evidence involving other goods may have fostered that misunderstanding. On that basis, the hypothesis that experimentally-observed WTP–WTA gaps in general are artifactual has seemed credible to many economists. Given that this disparity is one of the most widely-cited "anomalies" in individual decision-making, the truth or falsity of that hypothesis is a matter of considerable significance. Economists would be better able to reach informed judgments about this question if they also knew that the PZ elicitation procedure does not eliminate the disparity for lotteries. The main contribution of our paper is the presentation of that evidence.

[23] PZ's no-gap result for coffee mugs is replicated in an experiment reported by Stephanie Kovalchik et al. (2005).

The extent to which WTP–WTA disparities are artifactual needs to be reappraised in the light of this additional evidence. In the remainder of this paper, we discuss various other possibilities. We must emphasize that this discussion has a different status from the experimental results reported in Section III. Those results derive from experimental treatments that had been structured to investigate the extent of the WTP–WTA disparity for specific goods under specific elicitation procedures. Once we go beyond the questions that these experiments investigate, we move into a domain in which, it seems to us, the existing evidence base does not justify firm conclusions.

It seems clear that there is no WTP–WTA disparity for mugs under the PZ procedure. However, whether these disparities are caused by subject misconceptions remains an open question, particularly in the light of our controlled comparison between the PZ and KKT elicitation procedures, which found no significant differences in reported valuations.

The PZ procedure is primarily directed at correcting *a specific type* of misunderstanding by participants, namely misunderstanding of the BDM mechanism. PZ's investigative strategy seems to be guided by the hypothesis that WTP–WTA gaps are an artifact of elicitation mechanisms that either are not incentive compatible, or whose incentive-compatibility subjects do not fully understand.[24] However, while PZ's design goes to great lengths to correct this kind

[24] PZ emphasize four features of their procedure. First, it is incentive-compatible. Second, training "provides subjects with a basic understanding of the mechanism used to elicit valuations". Third, there are practice rounds in which participants "learn by gaining familiarity with the mechanism" and (in the paid practices) "learn about the intricacies of the elicitation mechanism and are given an opportunity to adjust nonoptimal strategies to maximize their payouts". Finally, decisions and payments are anonymous, to encourage participants to focus on their own rewards from the experiment (pp. 537–8). The common theme is that subjects are trained to maximize their reward from the experiment by reporting their "true" valuations.

of misunderstanding, it has other features which may dampen WTP–WTA disparities by reducing the salience of the distinction between buying and selling tasks. For example, PZ's instructions describe both buying and selling tasks as eliciting "offers" from the participant, rather than using terms such as "bids" and "asks" which might differentiate the tasks more. In the mug task, every participant is shown a mug; sellers are told that they own it, while buyers are told that they do not. But there is little else to flag up the difference between buying and selling, whereas other experiments draw more attention to this difference.

Subjects' perceptions of their reference state may be affected by factors such as ownership, physical possession of the object, whether or not endowments are determined at random, the wording of the task, and so on. For example, in Knetsch's (1989) classic investigation of willingness to exchange chocolates and mugs, goods are placed in front of the subjects who own them. Knetsch and Wei Kang Wong (2009) present experimental evidence which shows that subjects are reluctant to part with a mug or pen that they have in front of them, even if they do not own it, while the effect disappears if subjects *own* the object but do not have it with them at the moment of making their decisions. On the basis of such evidence, it is possible that WTP–WTA disparities may be attenuated if, as in PZ's design, *both* buyers and sellers have a mug in front of them. It seems that such effects are sensitive to subtle cues about reference points; but whatever these cues and their effects, there is no reason to assume that their being "turned off" is somehow the default state in transactions outside the laboratory.

A similar argument can be made about the effects of "training" and "practice" in the PZ design. While experience can be expected to reduce misunderstanding of experimental procedures, it may have other effects too. There is now considerable evidence that WTP–WTA disparities tend to decay as experimental subjects gain experience of buying and selling (e.g.

Jason Shogren et al., 1994; Loomes, Chris Starmer and Sugden, 2003), but it is not self-evident that the effect is mediated through increasing understanding of experimental procedures. One alternative hypothesis is that trading experience weakens an individual's perception of "not trading" as a salient reference point (Loomes et al., 2003). Another is that such experience reduces individuals' uncertainty about their own preferences; if there is loss aversion with respect to changes in utility, this will tend to reduce WTP–WTA disparities (Loomes et al, 2009). Some support for these hypotheses is given by John A. List's (2003) finding that, for a given set of experimental procedures, WTP–WTA disparities are smaller for subjects who have had more experience of buying and selling the relevant goods *outside* the experiment. If one is interested in the possibility that experience affects the extent of any anomaly, then what PZ call "paid practice" may be better interpreted as a treatment variable than as an essential control.

A further possibility is that, in PZ's implementation of their elicitation procedure and in our replication of this, the absence of WTP–WTA disparities for mugs is partly due to house money effects. In the original KKT experiment, and in PZ's replication of it, WTP–WTA disparities were found. In contrast, our controlled comparison found no differences between the KKT–ILS and PZ–ILS procedures, and no significant disparities for mugs in either case. Our experiment was not designed to investigate the effect of show-up fees, but we offer the conjecture that this combination of results may be due to the fact that, in the treatments which use the PZ procedure and in our KKT–ILS treatment, subjects who buy mugs can cover their expenditure from show-up fees (sometimes supplemented by receipts from sales of lotteries). In the original KKT experiment and in the KKT–PZ replication, there were no show-up fees and no opportunities to earn money prior to the mug tasks.

PZ argue that their data do not support the hypothesis of a house money effect. They report a regression analysis which shows that, in their Treatments 1 and 3, mug valuations were insensitive to earnings from preceding lottery tasks (pp. 541–2). We ran the same analysis on our Stage 1 data and found the same insensitivity. However, we are also able to compare valuations elicited using the PZ procedure in Stage 1, when the show-up fee was £3, with those in Stage 2, when it was £8.[25] Mean WTP was £1.86 in Stage 1 compared with £3.07 in Stage 2; the distributions of WTP valuations differ significantly ($p < 0.001$ in a two-tail Mann-Whitney test). On the other hand, there was no significant difference between the distributions of WTA valuations (the means were £2.21 in Stage 1 and £2.75 in Stage 2). These findings are consistent with the possibility that show-up fees induce a house money effect, while experimental earnings do not. Perhaps subjects assign show-up fees and experimental earnings to different "mental accounts" (Thaler, 1985) because show-up fees are interpreted as a budget that can be spent in the experiment.

While our results add to the evidence that (for whatever reason) the WTP–WTA disparity is absent for mugs under the PZ procedures, it seems to us no less significant a finding that the gap persists in the case of lotteries. The obvious inference to draw from this is that there is some systematic difference between mug and lottery tasks which generates this variability.

This cannot be attributed to variations in elicitation procedure, since the same procedure was applied in both cases. However, the mug and lottery tasks differ in some important respects. In the mug task, subjects report valuations for a *consumption good* to be obtained with *certainty*,

[25] This comparison is not as fully controlled as that between the KKT–ILS and PZ–ILS treatments. Subjects were recruited separately for the two Stages and (because of a rebranding exercise by the University of East Anglia) the mugs used in the later Stage displayed a different logo.

while in lottery tasks they value *sums of money* to be received with some *uncertainty*. This suggests that the difference between the mug and lottery data might be the result of either a difference between certainty and uncertainty or a difference between consumption goods and money. The results for the CG task (in the final column of panel A of Table 2), in which subjects value a *consumption good* to be obtained with some *uncertainty*, provide some suggestive evidence. Responses to this task show the same pattern as is found for the mug task: a small positive disparity (the ratio of mean WTA to mean WTP is 1.23) which is not statistically significant. This suggests that the relevant difference between the two types of task may be between money and consumption-good outcomes, rather than between uncertainty and certainty; but we recognize that the evidence base here is very small.

The difference between consumption goods and money may be significant because, in the PZ elicitation procedure, the *response mode* (that is, the form in which participants record their responses) is always the open-ended statement of a sum of money. In the lottery tasks, but not in mug or CG tasks, the response-mode units are also used in specifying the objects that are being valued. One possible effect of this is that lottery tasks may prompt respondents to use "anchoring" heuristics that are not applicable to the other tasks. In relation to lotteries that offer only two outcomes, one positive and relatively large and one zero (or close to zero), respondents might be expected to anchor on the former. This would induce a general tendency to over-value such lotteries. However, in order to induce a WTP–WTA gap, that effect would have to act disproportionately on selling tasks. As far as we know, existing theories of anchoring do not account for such an asymmetry.[26]

[26] Another mechanism through which the response mode might affect WTP and WTA valuations is analyzed in a theoretical paper by Andreas Lange and Anmol Ratan (in press). Following Kőszegi and Rabin (2006), Lange and

An alternative explanation of the difference between mug and lottery tasks is suggested by PZ in their footnote 15. Their conjecture is that tasks involving lotteries induce additional types of misconception, for which the PZ elicitation procedure does not control. In particular, subjects may have flawed understandings of the concepts of randomization and probability: "Experience seems to be necessary for subjects unfamiliar with random devices to incorporate true notions of randomization and the nature of probability" (pp. 539–540, note 15). In a private communication, PZ argue that the lottery data show various patterns that are inconsistent with both expected utility theory and EET, and that are indicative of misconceptions. Perhaps there is some way that misconceptions about probability interact with buying and selling so as to produce WTP–WTA gaps.

Such an account of the difference between mug and lottery tasks, like the others we have discussed, is a potentially credible ex post conjecture. However, it is a conjecture to the effect that, holding procedures constant, the extent of the WTP–WTA gap may vary according to the characteristics of the commodities being valued. Since the PZ procedure is designed to control for misconceptions about the elicitation mechanism, and since misconceptions about probability are related to the nature of the experimental good, this would be quite contrary to PZ's thesis, restated in their conclusion, that "[t]he differences reported in the literature reflect differences in experimental controls for misconceptions as opposed to differences in the nature of the commodity (e.g., candy, money, mugs, lotteries, etc.) under study" (p. 542).

Ratan model individuals' reference points as incorporating rational expectations of future exchanges and treat loss aversion as separable with respect to dimensions of consumption. In auctions in which individuals bid money to buy some item, loss aversion induces conservative bidding if the item is a non money commodity but aggressive bidding if it is an induced-value token, denominated in money. Since the usual WTP–WTA disparity corresponds with *conservative* bidding strategies, this model cannot explain the absence of a disparity for mugs in PZ's and our data.

PZ's strategy for investigating this thesis is to elicit WTP and WTA valuations using a procedure that incorporates all previously-used controls for elicitation-related misconceptions. In order to establish whether it is differences in procedures *as opposed to* differences between commodities that account for variability in the existing evidence, it would seem desirable to apply that procedure to more than one commodity. By applying the same elicitation procedure to mugs and lotteries, PZ's design makes this possible. If their lottery data were in any way contaminated, our additional controls have overcome that problem. Taking their full dataset in conjunction with ours, it is clear that when the PZ procedure is used, the WTP–WTA gap is absent for one type of good frequently used in experiments but is significant and persistent for another. By drawing attention to these data, we hope that we may have provided some stimulus for researchers in this field to investigate further the unresolved issues that we have highlighted.

REFERENCES

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential method." *Behavioral Science*, 9(3): 226–232.

Carrigan, Jay R., and Matthew C. Rousu. 2008. "Testing Whether Field Auction Experiments Are demand Revealing in Practice." *Journal of Agricultural and Resource Economics*, 33(2): 290–301.

Clark, Jeremy A. and Lana Friesen. 2008. "The Causes of Order Effects in Contingent Valuation Surveys: An Experimental Investigation." *Journal of Environmental Economics and Management*, 56(2): 195–206.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10(2): 171–178.

Flachaire, Emmanuel, and Guillaume Hollard. 2008. "Individual Sensitivity to Framing Effects." *Journal of Economic Behavior and Organization*, 67(1): 296–307.

Garnett, Stelle N. 2006. "The Neglected Political Economy of Eminent Domain." *Michigan Law Review*, 105: 101–150.

Hanemann, W. Michael. 1991. "Willingness to Pay and Willingness to Accept: How Much Can They Differ?" *American Economic Review*, 81(3): 635–647.

Horowitz, John K., and Kenneth E. McConnell. 2002. "A Review of WTA/WTP Studies." *Journal of Environmental Economics and Management*, 44(3): 426–447.

Huck, Steffen, Kirchsteiger, Georg and Jörg Oechssler. 2005. “Learning to Like What You Have – Explaining the Endowment Effect”. *The Economic Journal*, 115(505): 689–702.

Isoni, Andrea. 2009. “The Willingness-to-accept/Willingness-to-pay Disparity in Repeated Markets: Loss Aversion or ‘Bad-deal’ Aversion?” CSERGE Working Paper edm-2009-06, Centre for Social and Economic Research on the Global Environment, University of East Anglia.

Johnson, Eric J., Gerald Häubl, and Anat Keinan. 2007. “Aspects of Endowment: A Query Theory of Value Construction.” *Journal of Experimental Psychology – Learning, Memory, and Cognition*, 33(3): 461–474.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica*, 47(2): 263–291.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1990. “Experimental Tests of the Endowment Effect and the Coase Theorem.” *Journal of Political Economy* 98(6): 1325–1348.

Knetsch, Jack L. 1989. “The Endowment Effect and Evidence of Nonreversible Indifference Curves.” *American Economic Review*, 79(5): 1277–1284.

Knetsch, Jack L., and Wei-Kang Wong. 2009. “The Endowment Effect and the Reference State: Evidence and Manipulations.” *Journal of Economic Behavior and Organization*, 71(2): 407–413.

Kolstad, Charles D., and Rolando M. Guzman. 1999. “Information and the Divergence between Willingness to Accept and Willingness to Pay.” *Journal of Environmental Economics and Management*, 38(1): 66–80.

Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121(4): 1133–1166.

Kovalchik, Stephanie, Colin F. Camerer, David M. Grether, Charles R. Plott, and John M. Allman. 2005. "Aging and Decision Making: A Comparison Between Neurologically Healthy Elderly and Young Individuals." *Journal of Economic Behavior and Organization*, 58(1): 79–94.

Lange, Andreas, and Anmol Ratan. Forthcoming. "Multi-dimensional Reference-dependent Preferences in Sealed-bid Auctions: How (Most) Laboratory Experiments Differ from the Field." *Games and Economic Behavior*.

List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, 118(1): 41–71.

Loomes, Graham, Shepley Orr, and Robert Sugden. 2009. "Taste Uncertainty and Status Quo Effects in Consumer Choice." *Journal of Risk and Uncertainty* 39(2): 113–135.

Loomes, Graham, Chris Starmer, and Robert Sugden. 2003. "Do Anomalies Disappear in Repeated Markets?" *Economic Journal*, 113(486): C153–C166.

Mandler, Michael. 2004. "Status Quo Maintenance Reconsidered: Changing or Incomplete Preferences?" *Economic Journal*, 114(499): F518–F535.

Plott, Charles R., and Kathryn Zeiler. 2002. "The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect" and Experimental Procedures for Eliciting Valuations." Unpublished paper, California Institute of Technology.

Plott, Charles R., and Kathryn Zeiler. 2005. "The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review*, 95(3): 530–545.

Plott, Charles R., and Kathryn Zeiler. 2007. "Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory?" *American Economic Review*, 97(4): 1449–1466.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292.

Rachlinsky, Jeffrey J. 2006. "Bottom-up versus Top-down Lawmaking." *The University of Chicago Law Review*, 73(3): 933–964.

Shogren, Jason, Seung Y. Shin, Dermot Hayes, and James B. Kliebenstein. 1994. "Resolving Differences in Willingness to Pay and Willingness to Accept." *American Economic Review*, 84(1): 255–270.

Sugden, Robert. 2003. "Reference-Dependent Subjective Expected Utility." *Journal of Economic Theory*, 111(2): 172–91.

Thaler, Richard. 1980. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behavior and Organization*, 1(1): 39–60.

Thaler, Richard. 1985. "Mental Accounting and Consumer Choice." *Marketing Science*, 4(3): 199–214.

Thaler, Richard, and Eric Johnson. 1990. "Gambling With the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choices." *Management Science*, 36(6): 643–660.

Tsur, Matan. 2008. "The Selectivity Effect of Past Experience on Purchasing Decisions: Implications for the WTA–WTP Disparity." *Journal of Economic Psychology*, 29(5): 739–746.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics*, 106(4): 1039–1061.

Willig, Robert D. 1976 "Consumer's Surplus without Apology." *American Economic Review*, 66(4): 589–597.

Table 1 - Lottery parameters

	Val. Type	Lott. No.	Replication Lottery	Plott & Zeiler (2005)	
				Lottery A	Lottery B
Small-stake lotteries	WTA	1	(£0.20, 0.5; £0.20, 0.5)	($0.20, 0.5; $0.20, 0.5)	($0.20, 0.5; $0.20, 0.5)
		2	(£0.30, 0.5; £0.30, 0.5)	($0.35, 0.5; $0.35, 0.5)	($0.35, 0.5; $0.35, 0.5)
		3	(£0.70, 0.5; £0, 0.5)	($0.70, 0.3; $-0.20, 0.7)	($-0.20, 0.3; $0.70, 0.7)
	WTP	4	(£0.30, 0.5; £0.30, 0.5)	($0.30, 0.5; $0.30, 0.5)	($0.30, 0.5; $0.30, 0.5)
		5	(£0.40, 0.5; £0.40, 0.5)	($0.45, 0.5; $0.45, 0.5)	($0.45, 0.5; $0.45, 0.5)
		6	(£0.80, 0.5; £0.10, 0.5)	($0.80, 0.3; $-0.10, 0.7)	($-0.10, 0.3; $0.80, 0.7)
Large-stake lotteries	WTA	7	(£3, 0.7; £0, 0.3)	($7, 0.7; $0, 0.3)	($0, 0.7; $7, 0.3)
		8	(£2, 0.4; £0, 0.6)	($5, 0.4; $0, 0.6)	($0, 0.4; $5, 0.6)
		9	(£2.50, 0.5; £0.50, 0.5)	($8, 0.5; $-4, 0.5)	($-4, 0.5; $8, 0.5)
		10	(£4, 0.3; £0, 0.7)	($10, 0.3; $0, 0.7)	($0, 0.3; $10, 0.7)
	WTP	11	(£4, 0.7; £1, 0.3)	($8, 0.7; $1, 0.3)	($1, 0.7; $8, 0.3)
		12	(£3, 0.4; £1, 0.6)	($6, 0.4; $1, 0.6)	($1, 0.4; $6, 0.6)
		13	(£3.50, 0.5; £1.50, 0.5)	($9, 0.5; $-3, 0.5)	($-3, 0.5; $9, 0.5)
		14	(£5, 0.3; £1, 0.7)	($11, 0.3; $1, 0.7)	($1, 0.3; $11, 0.7)

Table 2 - Experimental Results

A) Replication Experiment – Stage 1

WTA valuation	**L1**	**L2**	**L3**	**L7**	**L8**	**L9**	**L10**	**Mug**	**CG**
n	100	100	100	100	100	100	100	51	49
EV	0.20	0.30	0.35	2.10	0.80	1.50	1.20		
Mean	0.23	0.31	0.38	2.16	0.94	1.40	1.57	2.21	2.15
Median	0.20	0.30	0.30	2.10	0.85	1.50	1.20	2.00	1.50
Standard Deviation	0.29	0.14	0.53	0.72	0.43	0.50	0.96	1.80	2.09
Mean/EV	1.17	1.03	1.09	1.03	1.18	0.93	1.31		

WTP valuation	**L4**	**L5**	**L6**	**L11**	**L12**	**L13**	**L14**	**Mug**	**CG**
n	100	100	100	100	100	100	100	49	51
EV	0.30	0.40	0.45	3.10	1.80	2.50	2.20		
Mean	0.29	0.43	0.35	2.49	1.57	2.31	2.24	1.86	1.75
Median	0.30	0.40	0.30	2.50	1.50	2.25	2.00	1.80	1.00
Standard Deviation	0.07	0.17	0.26	1.11	0.52	0.64	1.12	1.29	1.68
Mean/EV	0.95	1.07	0.78	0.80	0.87	0.92	1.02		

WTA/WTP[a]	**L1/L4**	**L2/L5**	**L3/L6**	**L7/L11**	**L8/L12**	**L9/L13**	**L10/L14**	**Mug**	**CG**
Mean	1.18	1.02	2.19	1.53	1.37	1.11	1.46	1.19	1.23
Median	1.00	1.00	1.33	1.26	1.16	1.00	1.11	1.11	1.50
Significance[b]	n/a	n/a	***	***	***		***		

Table 2 *(continued)*

B) Replication Experiment – Stage 2: Comparison between PZ-ILS and KKT-ILS Treatments (mugs only)

WTA valuation[c]	**PZ-ILS**	**KKT-ILS**
n	33	39
Mean	2.75	2.85
Median	3.00	2.75
Standard Deviation	1.76	1.86

WTP valuation[c]	**PZ-ILS**	**KKT-ILS**
n	33	39
Mean	3.07	2.96
Median	2.50	2.25
Standard Deviation	1.53	2.40

WTA/WTP[a]	**PZ-ILS**	**KKT-ILS**
Mean	0.90	0.96
Median	1.20	1.22
Significance[b]		

Table 2 *(continued)*

C) PZ Experiment – A Lotteries (Treatments 1, 2 and 3 pooled)

WTA valuation	**L1**	**L2**	**L3**	**L7**	**L8**	**L9**	**L10**
n	36	36	36	36	36	36	36
EV	0.20	0.35	0.07	4.90	2.00	2.00	3.00
Mean	0.20	0.35	0.20	4.81	2.68	2.87	3.86
Median	0.20	0.35	0.10	4.95	2.15	2.00	3.00
Standard Deviation	0.02	0.01	0.21	1.48	1.08	1.88	2.53
Mean/EV	0.99	1.00	2.87	0.98	1.34	1.43	1.29

WTP valuation	**L4**	**L5**	**L6**	**L11**	**L12**	**L13**	**L14**
n	36	36	36	36	36	36	36
EV	0.30	0.45	0.17	5.90	3.00	3.00	4.00
Mean	0.30	0.45	0.23	4.86	2.63	3.45	4.24
Median	0.30	0.45	0.18	5.15	2.90	3.00	4.00
Standard Deviation	0.01	0.02	0.20	1.59	0.96	2.04	2.58
Mean/EV	0.99	1.01	1.33	0.82	0.88	1.15	1.06

WTA/WTP[a]	**L1/L4**	**L2/L5**	**L3/L6**	**L7/L11**	**L8/L12**	**L9/L13**	**L10/L14**
Mean	1.00	0.99	1.97	1.47	1.66	1.38	1.46
Median	1.00	1.00	1.23	1.08	1.23	1.00	1.01
Significance[b]	n/a	n/a	***	***	***	**	**

Table 2 *(continued)*

D) PZ Experiment – B Lotteries (Treatments 1, 2 and 3 pooled)

WTA valuation	**L1**	**L2**	**L3**	**L7**	**L8**	**L9**	**L10**
n	38	38	38	38	38	38	38
EV	0.20	0.35	0.43	2.10	3.00	2.00	7.00
Mean	0.20	0.35	0.44	2.67	2.80	2.69	6.78
Median	0.20	0.35	0.45	2.10	3.00	2.00	7.00
Standard Deviation	0.00	0.01	0.17	1.56	0.99	1.81	1.70
Mean/EV	1.00	1.00	1.01	1.27	0.93	1.34	0.97

WTP valuation	**L4**	**L5**	**L6**	**L11**	**L12**	**L13**	**L14**
n	38	38	38	38	38	38	38
EV	0.30	0.45	0.53	3.10	4.00	3.00	8.00
Mean	0.30	0.45	0.49	2.41	3.10	2.67	7.03
Median	0.30	0.45	0.50	2.48	3.00	3.00	7.41
Standard Deviation	0.00	0.01	0.18	0.76	1.07	1.24	2.11
Mean/EV	1.00	0.99	0.92	0.78	0.78	0.89	0.88

WTA/WTP[a]	**L1/L4**	**L2/L5**	**L3/L6**	**L7/L11**	**L8/L12**	**L9/L13**	**L10/L14**
Mean	1.00	1.00	1.13	1.67	1.34	1.97	1.20
Median	1.00	1.00	1.07	1.36	1.20	1.34	1.08
Significance[b]	n/a	n/a	**	***	***	***	**

Table 2 *(continued)*

E) PZ experiment – Mugs

WTA valuation	**Pooled**	**Mugs Last**	**Mugs First**
n	38	24	14
Mean	5.56	5.48	5.71
Median	5.00	5.00	5.10
Standard Deviation	3.58	3.40	4.00
WTP valuation	**Pooled**	**Mugs Last**	**Mugs First**
n	36	24	12
Mean	6.62	5.99	7.88
Median	6.00	6.00	6.50
Standard Deviation	4.20	2.90	6.00
WTA/WTP[a]	**Pooled**	**Mugs Last**	**Mugs First**
Mean	0.84	0.92	0.72
Median	0.83	0.83	0.78
Significance[b]			

a – Ratio is computed as (WTA + c)/WTP for lotteries, while for the mug and CG it is the ratio of means and medians respectively. The constant c is £0.10 ($0.10) for small-stake lotteries (1-6) and £1 ($1) for high-stake lotteries (7-14).

b – Test based on signed ranks for lotteries and for sum or ranks for mug and CG. Significance level (1-tail): * = 10%, ** = 5%, *** = 1%. Test not reported for certainties.

c – No statistically significant difference between distributions of valuations in PZ-ILS and KKT-ILS treatments (two-tail rank sum test).