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

Andrea Isoni, Graham Loomes and Robert Sugden*

23 January 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: Centre for Behavioural and Experimental Social Science and School of Economics, University of East Anglia, Norwich (UK), NR4 7TJ (email: g.loomes@uea.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.

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

Abstract

Charles R. Plott and Kathryn Zeiler (2005) report the results of a four-part experiment in support of the claim that the commonly observed willingness-to-pay/willingness-to-accept disparity is due to subjects' misconceptions about the experimental procedures used to elicit valuations. We express doubts about whether Plott and Zeiler's data justify this claim, and report the results of an experiment aimed at resolving these doubts by replicating their procedures as closely as possible. We conclude that Plott and Zeiler's claim is supported neither by the full data set generated by their experiments nor by our replication.

Keywords: WTP-WTA disparity, endowment effect, misconceptions, reference-dependence.

JEL classification: C91, D11.

During the last four decades, numerous contingent valuation and experimental studies have reported discrepancies between willingness-to-pay (WTP) and willingness-to-accept (WTA) measures of value that are far larger than those predicted by 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). These findings have generally been interpreted as evidence of a systematic asymmetry between individuals' attitudes to gains and losses relative to some reference point. 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 Köszegi and Matthew Rabin, 2006; Graham Loomes, Shepley Orr and Sugden, 2007).

In a recent article, Plott and Zeiler (2005) – henceforth PZ – oppose this widely-accepted interpretation of the evidence, hypothesising that 'observed WTP-WTA gaps do not reflect a fundamental feature of human preferences', but instead are 'symptomatic of subjects' misconceptions about the nature of the experimental task' in which valuations are elicited; it is their belief that 'differences [in the extent of the gaps] 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). They claim that their hypothesis is supported by the results of three new experimental treatments reported in their paper,[1] contrasted with the results of a replication of a previous experiment which found a WTP-WTA disparity.

PZ review a number of experimental investigations of WTP and WTA for a range of different 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.[2] Noting that different experiments have used different procedures to try to reduce subjects' misconceptions, but acknowledging that no comprehensive theory of misconceptions exists, PZ pose the following 'main research question': 'If we design an

[1] PZ refer to these as 'Experiments' 1, 2 and 3; for clarity, we will use the term 'Treatments'.

[2] In an earlier working paper (Plott and Zeiler, 2003, Table 5), PZ classify these experiments by the commodities for which valuations were elicited. The commodities include 'lotteries for goods' and 'lottery tickets'.

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. 531-532)

PZ present their experimental design as one in which 'subject misconceptions are completely controlled by incorporating the union of procedures found in the literature' (pp. 531–532). Employing this *revealed theory methodology*, they claim to establish the following 'striking result': 'When 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, we observe no WTP-WTA gap' (pp. 531–532). We will call this PZ's *no-gap claim*. PZ do not specify the domain of this claim, but in their Abstract, they 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' (p. 530). The clear implication is that the procedures deployed by PZ eliminate the WTP-WTA gap for both lotteries and mugs. However, we question whether this implication is supported by their own data, and we report the results of an experiment which aims to resolve any doubts by closely replicating PZ's design.[3]

First we note that PZ's conclusions are based on only a small fraction of the data generated by their experiment. In each of their three new treatments, each participant faced 15 incentive-compatible tasks, 14 eliciting WTP or WTA for lotteries and one eliciting WTP or WTA for mugs.[4] But PZ report only the results of the mug tasks, treating the lottery tasks as 'paid practice' (if coming first as in Treatments 1 and 3) or as irrelevant (if coming second as in Treatment 2). However, as we shall see, the lottery tasks are set up in a way which makes them well-suited to test for WTP-WTA disparities. Given that lotteries *are* commonly used in

[3] As well as investigating whether WTP-WTA gaps are explained by subject misperceptions, PZ address the separate issue of whether such gaps are explained by 'endowment effect theory' (their term for a body of ideas 'associated with' prospect theory [p. 531]. In this paper, we are concerned only with the no-gap claim. We do not here make any judgement about the empirical validity of prospect theory, and testing that is not any part of our objective.

[4] In their paper, PZ say very little about the lottery tasks, apart from noting some 'speculations and conjectures' based on their interpretation of the data (pp. 539-540, note 15). In the online Appendix, they describe the lottery tasks in more detail, but not the data that they generated (http://www.e-aer.org/data/june05_app_plott.pdf). We thank them for giving us access to these data.

experiments, and that PZ's review ranged over previous experiments involving both mugs and lotteries, it seems natural to examine the impact of their 'controls' on the lottery tasks as well as the mug tasks.[5]

PZ decline to use their lottery data for testing hypotheses, on the grounds that these data were 'contaminated by a design that was developed only for training'. They refer to two forms of 'contamination'. First, lottery selling tasks preceded lottery buying tasks. Second (and presumably because lottery tasks were sometimes used as 'paid practices'): 'Mistake 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).

However, we suggest that PZ's design, as described in the experimental materials and formal instructions,[6] *does* generate useable lottery data. Certainly, the order in which buying and selling tasks were presented does not rule out tests for WTP-WTA disparities. It is more difficult to judge the status of PZ's concerns about contamination from 'mistake corrections', since these involve informal features of the procedures, not described in the formal instructions, making it difficult for anyone else to assess the extent of any potential contamination.

To overcome this latter problem, we replicate PZ's experiment as closely as possible, while ensuring that *none* of the 15 paid tasks is contaminated by informal training procedures. A replication of this kind allows us both to check the robustness of PZ's no-gap claim with respect to mugs, and to investigate whether it holds for lotteries. Our results replicate PZ's finding of no significant WTP-WTA disparity for mugs. They also replicate the salient features of their unpublished data, namely a significant disparity for lotteries which shows no tendency to decay as participants gain experience of the experimental procedures. We conclude that PZ's unqualified no-gap claim is not supported. Moreover, by disregarding their lottery data, they appear to have missed an interesting and potentially important finding concerning differences between mugs and the binary lotteries so often used in economic experiments.

---

[5] Lottery tasks are also relevant for PZ's objective of testing 'endowment effect theory'. Although the original form of prospect theory required reference points to be certainties, and so could not make predictions about WTA for lotteries (Kahneman and Tversky, 1979), later versions of the theory predict WTP/WTA disparities for lotteries (Sugden, 2003; Köszegi and Rabin, 2006).

[6] The formal instructions are reproduced in PZ's online Appendix: see note 3 above.

In addition to the three treatments using their own procedures, PZ report a replication of a design used by Kahneman, Jack L. Knetsch and Richard Thaler (1990) – henceforth KKT – to elicit valuations for mugs. Using this design, which they treat as representative of 'procedures commonly used in studies that report observed gaps', PZ replicate KKT's finding of a significant WTP-WTA disparity. Comparing this result with those of their other three treatments, PZ conclude that WTP-WTA disparities can be 'turned on and off' by using procedures which differ in the extent to which they control for misconceptions (p. 542). In order to ensure a full replication of PZ's entire experiment, we ran a second stage of our investigation, comparing the KKT and PZ designs.[7] By contrast with PZ, who used different subject pools and different goods in the various treatments,[8] we implemented a controlled comparison in which subjects were randomised between treatments, with the same good being used throughout. Moreover, whereas PZ gave a show-up fee in their own treatment 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 claim, but it raises questions about how far the WTP-WTA disparity is attributable to misconceptions.

I. Plott and Zeiler's design

In this Section, we briefly review PZ's design. We focus on those properties of the lottery tasks that allow them to be used to investigate WTP-WTA disparities. We also point to some features of the procedures (used for both lottery and mug tasks) which may make the distinction between WTP and WTA tasks less salient to respondents than in some other experiments. These features might be expected to dampen WTP-WTA disparities, but it is not obvious that their inclusion in the design increases the extent of control for misconceptions. (For details of other aspects of the design, see Plott and Zeiler, 2003, 2005.)

---

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

[8] PZ's Treatments 1 and 2 used students at the University of Southern California Law School; Treatment 3 used students at Pasadena City College; and the KKT replication experiment used students at CalTech, with the last of these treatments involving a different mug than the other three.

The overall structure of the experiment 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 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 organised 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 realisation 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, 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 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*

[9] The purpose of the lottery tasks in Treatment 2 is not clear. According to PZ, these (rather expensive) tasks were 'developed only for training and not for the purpose of measuring a gap' (p. 539, note 15); but they appear *after* the mug task.

[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*.

the corresponding WTA valuations (that is, when participants' accumulated earnings are expected to be higher), any wealth effects will tend to make WTP higher that it would be under the 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.

It might seem that the presentation of WTA and WTP tasks in blocks of three (small-stake lotteries) or four (large-stake lotteries), with WTA tasks preceding WTP ones, is less than ideal. 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. And the use of blocks of same-type tasks has compensating advantages. A design in which participants constantly switched between buying and selling tasks would be liable to cause confusion, undermining the aim of eliminating misconceptions. In addition, as PZ explain, having a block of WTA tasks at the beginning of the experiment allows participants to accumulate earnings to spend in WTP tasks.

The tests that can be carried out using the lottery data are within-subject, whereas the mug tasks generate data for between-subject tests. The latter tests require a split sample, and so are much less powerful. The small numbers of participants in the PZ experiment (31, 26 and 17 in Treatments 1, 2 and 3 respectively) make considerations of statistical power particularly relevant.[11]

Overall, then, the lottery tasks appear no less well-designed than the mug tasks for testing hypotheses about WTP-WTA disparities, given the constraints of sample size. So if one considers only the design of the experiment as it is represented in the experimental materials and formal instructions, there seems to be no prima facie reason to treat the mug task as the 'real' test

[11] In fact, the unusually low WTA/WTP ratio (0.84) that PZ find for mugs is too low to have been generated by plausible sampling error from a population in which the true ratio is much above unity. But, from an ex ante viewpoint, there are strong reasons for preferring within-subject to between-subject tests when the sample size is small.

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: the statement that one paid task is a 'practice' for another has no clear meaning beyond the fact that the former precedes the latter. This point applies particularly to the large-stake lotteries. Arguably, the very low payoffs in the small-stake lotteries, and the fact that four out the six small-stake tasks involved degenerate lotteries, justify their being interpreted as 'training' rather than 'real' tasks. This interpretation is in fact consistent with PZ's instructions, which seem to imply that participants should treat the *small-stake* lotteries as practices, not just for the mug tasks, but also for the large-stake lottery tasks. In PZ's words, 'subjects were told that the lotteries would increase in magnitude, but *the first few rounds* allowed for additional (but paid) practice' (p. 538, emphasis added).[12]

Any hypothesis about the effects of paid practice can be more usefully reformulated as a hypothesis about the effects of *experience* – that is, of repeatedly facing paid tasks of a given type. Thus, the most efficient way to use the data generated by PZ's design is to treat *both* lottery *and* mug tasks as tests for disparities, and to investigate whether the size and frequency of disparities change as participants gain experience. By allowing their lottery data to be contaminated by 'mistake corrections' and 'public answers to questions', PZ discard a large amount of potentially valuable data for little obvious gain.

Some of PZ's procedures are directed at removing possible misconceptions about the workings of the BDM mechanism: in particular, they aim to disabuse any participant who mistakenly thinks she can benefit by misrepresenting her true valuations. In addition, the presentation of WTP and WTA tasks seems liable to minimise differences between the two as perceived by the participants. 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

[12] PZ's actual instructions ended as follows: 'Before we begin, note that 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.'

flag up the difference between buying and selling, whereas many WTP-WTA experiments draw more attention to this difference.[13]

PZ may consider the minimisation of differences between WTA and WTP tasks to be an element of the set of 'controls' implemented in their design.[14] But, with equal legitimacy, one might argue that making the buy/sell distinction salient in a laboratory experiment simulates cues which are present in naturally-occurring economic environments, and that the absence of such cues reduces the external validity of experimental results. Ultimately, there is no unambiguously correct answer to the question 'What is the most controlled design for testing for WTP-WTA disparities?' We suggest that the PZ design is best understood as a stress test: it investigates whether disparities occur in an environment that is particularly unfavourable to them. Arguably, this makes the occurrence and persistence of such disparities in the case of lotteries all the more significant.

II. Our design

We begin by describing Stage 1 of our investigations. In this stage, our aim was to replicate PZ's procedures, while ensuring that the lottery data were not contaminated. Most of the differences between the original design and the replication are adaptations necessary for a computerised implementation, rather than PZ's pen-and-paper methods. We chose to use computers not only to simplify the organisation of the experiment, but also in the interests of replicability, by making the interface between participant and experiment as far as possible pre-scripted and so open to subsequent inspection.

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

[13] 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) provide evidence that the location of the good during an elicitation experiment has a significant effect on valuations. We discuss this issue further in Section IV.

[14] In another paper, Plott and Zeiler (2007, p. 1455) argue that procedures which 'emphasize entitlement' might be perceived as ' inadvertently signalling that the endowed good is more valuable than the alternate good' and/or that 'the "correct" choice is to keep the endowed good'.

entirely standard visual representation of lotteries.[15] They were read out by an experimenter while participants followed the text in printed form. The full text of the instructions can be found in the Appendix [intended only for online publication].

In the 'worked example' 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 coloured envelope randomly selected from a set of 80.[16] 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 prizes, the computer displayed an error message explaining why the value was not optimal given the reward mechanism. Before proceeding, the experimenter clarified any doubts regarding the message on the screen 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

[15] We thank Kathryn Zeiler for her assistance in the preparation of the experimental instructions.

[16] Two sets of coloured 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.

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 *[deleted for anonymity]* 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.[17] Participants who reported 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.[18]

---

[17] 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.

[18] We thank Kathryn Zeiler and [*deleted for anonymity*] for their suggestions about the design of this procedure.

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. PZ treat their KKT replication as a benchmark against which to measure the effectiveness of their controls for misconceptions. When repeating this part of the PZ experiment, we tried to make the two treatments as comparable as possible, by allowing them to differ 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 and KKT treatments. Each treatment elicited WTP and WTA for a mug, the same type of mug in both treatments.

The PZ treatment was essentially the same as in Stage 1, 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.[19] In this respect, the status of mug tasks in our PZ 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 computerisation of the design, we had substituted an on-screen 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 PZ paper, the KKT treatment elicited hypothetical WTP or WTA valuations for two fixed-value tokens (the same tokens as in the PZ treatment), prior to the mug task. Buyers

[19] 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.

and sellers sat in adjacent seats; a mug was placed in front of each seller, and buyers could inspect this.[20] In the interests of greater comparability with our PZ replication, we made two changes. First, our implementation was computerised. Second, we paid the same £8.00 show-up fee as in the PZ treatment. This latter change was introduced in order to control for a potentially confounding factor. We cannot rule out the possibility that a show-up fee may increase WTP through a house money effect (Richard Thaler and Eric Johnson, 1990), thereby reducing or eliminating any WTP-WTA gap that would otherwise be observed. If a show-up fee were paid in the PZ treatment but not in the KKT treatment, a comparison between the two treatments would be unable to disentangle house money effects from any effects due to PZ's controls for misconceptions.[21]

III. Results

Both stages of the experimental investigation were conducted at the *[name of laboratory deleted for anonymity]* using the Zurich Toolbox 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 row 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). For each of the two tasks, the table shows: the number of observations; (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 'standardised WTA/WTP' statistics. For the lottery tasks, standardised ratios are defined as $[\text{WTA}(L) + c] / \text{WTP}(K)$;[22] the statistics reported are the means and medians of the within-

[20] The full instructions of the KKT treatment are reported in the Appendix.

[21] PZ discuss the role of house money effects in their experiment, but only with respect to the earnings from the lottery tasks. They report that 'none of the variation in mug valuations is explained by variation in income earned during the practice rounds' (p. 542). However, since their analysis uses only data from their own treatments, in which every participant was paid the same show-up fee, it cannot address the potential confounding role of show-up fees in their comparison with the KKT design.

[22] Relative to the obvious alternative, namely $\text{WTA}(L) / [\text{WTP}(K) - c]$, this definition gives lower values and is compatible with observations for which $\text{WTP}(K) \leq c$.

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 column reports the result of a test of the hypothesis that, after standardisation, WTA is greater than WTP.[23] For lottery tasks, the significance level reported in the last column 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 (1, 2, 4 and 5). 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 column of panel A of Table 2, WTA significantly exceeds WTP (always at the 1 percent level) in four of the five possible comparisons.[24] 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, standardised 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

[23] 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 WTA-WTP 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 WTA/WTP disparities.

[24] 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 (in retrospect, at least) that PZ's concern about the possible contamination of their lottery data was unnecessary.

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 standardised 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 is variation over time, but looking at the data as a whole, this variation 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 results reported by PZ in support of their no-gap claim. 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 procedures are 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 randomised between the PZ and KKT 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. Conclusions

Our primary conclusion is that neither PZ's data nor ours support PZ's no-gap claim, if that claim applies not only to mugs but also to lotteries. 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 experimental procedures, 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 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 rationalise 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.[25]

Even 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 systematic disparity in the PZ lottery data could be an artefact of contamination from this source. But 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 under the PZ procedures the WTP-WTA disparity *is absent for mugs but occurs and persists for money lotteries*.

With respect to mugs, this claim is supported by all three of PZ's own treatments, by our Stage 1 and Stage 2 replications, and by an experiment reported by Stephanie Kovalchik et al. (2005) which found no significant disparity when using the PZ procedures to elicit WTP and WTA valuations of coffee mugs.

---

[25] PZ have suggested that WTP-WTA gaps for lotteries might be the result of other, lottery-specific miconceptions, not controlled for in their design. We comment on this possibility at the end of this Section.

From this evidence, however, we are not entitled to conclude that WTP-WTA disparities for mugs are a laboratory artefact, or that they are caused by subject misconceptions. The PZ procedures are primarily directed at correcting *a specific type* of misunderstanding by participants – misunderstanding of the BDM mechanism – which, if not corrected, may be liable to increase WTP-WTA disparities. However, as we noted in Section I, there are other features of their design which may dampen disparities by reducing the salience of the distinction between buying and selling tasks. One crucial aspect in this distinction is subjects' perception of their reference state. This 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. Knetsch and Wong (2009) present experimental evidence which shows that subjects are relucatant 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 under the PZ procedures, both buyers and sellers have a mug in front of them. It seems that such effects are sensitive to subtle cues about reference points, which cannot be expected to have exactly the same effects across laboratories, goods and subject pools. Whatever these cues and their effects, there is no reason to assume that they occur only in the laboratory and that their being 'turned off' is somehow the default 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. An alternative hypothesis is that trading experience weakens an individual's perception of 'not trading' as a salient reference point (Loomes et al., 2003). Some support for this hypothesis 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, when the PZ procedures are used, the absence of WTP-WTA disparities for mugs is partly due to house money effects.[26] 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 and PZ procedures, and no significant disparities for mugs in either case. One explanation for this otherwise surprising combination of results is that WTP-WTA disparities are attenuated when (as in the PZ procedures and in our KKT treatment) subjects who buy mugs can cover their expenditure from show-up fees (sometimes supplemented by receipts from sales of lotteries). To the extent that house money effects are artefacts of the laboratory environment, the *absence* of disparities might be regarded as artefactual.

However, the external validity of the different experimental procedures is not our primary focus. While our results add to the evidence that (for whatever reason) the disparity is absent for mugs under the PZ procedures, it seems to us no less significant a finding (supported by PZ's data as well as by ours) that the same is *not* true for lotteries. The most natural inference to draw from this is that there is some structural difference between mug and lottery valuation tasks, such that when the same experimental procedures are applied to both, WTP-WTA disparities can persist in lottery valuations but not in mug valuations.

What is the relevant difference between the mug and lottery tasks? The results for the CG task (in the final row of panel B of Table 2) 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. We offer the following conjecture. In all PZ tasks, the *response mode* (that is, the form in which participants record their responses) is 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. Thus,

[26] These effects might also help to explain the relatively low WTA/WTP ratios found in our (and PZ's) lottery tasks.

lottery tasks may prompt respondents to use 'anchoring' heuristics that are not applicable to the other tasks, and these may in some way contribute to WTP-WTA disparities.[27]

An alternative explanation of the difference between mug and lottery tasks is suggested by PZ in the footnote in which they explain why they do not report their lottery data. The suggestion is that tasks involving lotteries with monetary outcomes induce additional types of misconception, for which the PZ procedures do not control (p. 540). Of course, that could be so. Whatever anomalies one finds in an experiment, it is always possible to postulate some 'subject misconception' that might have caused it. That response is always available because, as PZ note at the start of their paper, no well-developed theory of misconceptions exist. But recall that PZ's revealed theory methodology was chosen to avoid exactly that problem. The problem was to be avoided by implementing the union of controls used in previous WTP-WTA experiments (including lottery experiments). Given that this methodology has been used, it is important to record the results that it has generated.

Recall that the revealed theory methodology was chosen to answer the 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?' PZ's data and ours both suggest that the answer is: 'For mugs, no; for lotteries, yes'.

[27] Loomes, Starmer and Sugden (2007) suggest that a mechanism of this kind might explain the finding (reported in their paper) that the preference reversal phenomenon does not decay as experimental subjects gain experience.

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.

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

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

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*, forthcoming.

Köszegi, 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

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. 2007. "Preference Uncertainty and Status Quo Effects in Consumer Choice." University of East Anglia.

Loomes, Graham, Chris Starmer, and Robert Sugden. 2003. "Do Anomalies Disappear in Repeated Markets?" *Economic Journal*, 113: C153-66.

Loomes, Graham, Chris Starmer, and Robert Sugden. 2007. "Preference Reversals and Disparities between Willingness to Pay and Willingness to Accept in Repeated Markets." University of East Anglia.

Plott, Charles R., and Kathryn Zeiler. 2003. "The Willingness to Pay/illingness to Accept Gap, the 'Endowment Effect', Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." Social Science Working Paper 1132, California Institute of Technology.

Plott, Charles R., and Kathryn Zeiler. 2003. "The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect", Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." 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.

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, 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.

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

Table 1 - Lottery tickets

	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) Stage 1 – Replication experiment

WTA valuation							WTP valuation							WTA /WTP[a]		Sign.[b]
Lott. No.	Obs.	EV	Mean	Median	Std. Dev.	Mean / EV	Lott. No.	Obs.	EV	Mean	Median	Std. Dev.	Mean / EV	Mean	Median	
1	100	0.20	0.23	0.20	0.29	1.17	4	100	0.30	0.29	0.30	0.07	0.95	1.18	1.00	n/a
2	100	0.30	0.31	0.30	0.14	1.03	5	100	0.40	0.43	0.40	0.17	1.07	1.02	1.00	n/a
3	100	0.35	0.38	0.30	0.53	1.09	6	100	0.45	0.35	0.30	0.26	0.78	2.19	1.33	***
7	100	2.10	2.16	2.10	0.72	1.03	11	100	3.10	2.49	2.50	1.11	0.80	1.53	1.26	***
8	100	0.80	0.94	0.85	0.43	1.18	12	100	1.80	1.57	1.50	0.52	0.87	1.37	1.16	***
9	100	1.50	1.40	1.50	0.50	0.93	13	100	2.50	2.31	2.25	0.64	0.92	1.11	1.00	
10	100	1.20	1.57	1.20	0.96	1.31	14	100	2.20	2.24	2.00	1.12	1.02	1.46	1.11	***
Mug	51		2.21	2.00	1.80		Mug	49		1.86	1.80	1.29		1.19	1.11	
CG	49		2.15	1.50	2.09		CG	51		1.75	1.00	1.68		1.23	1.50	

B) Stage 2 – PZ-KKT comparison (mugs only)

WTA valuation[c]					WTP valuation[c]				WTA /WTP[a]		Sign.[b]
	Obs.	Mean	Median	Std. Dev.	Lott. No.	Obs.	Mean	Median	Std. Dev.	Mean	Median	
PZ	33	2.75	3.00	1.76	PZ	33	3.07	2.50	1.53	0.90	1.20	
KKT	39	2.85	2.75	1.86	KKT	39	2.96	2.25	2.40	0.96	1.22	

C) PZ experiment – A lotteries (Treatments 1, 2 and 3)

WTA valuation							WTP valuation							WTA /WTP[a]		Sign.[b]
Lott. No.	Obs.	EV	Mean	Median	Std. Dev.	Mean / EV	Lott. No.	Obs.	EV	Mean	Median	Std. Dev.	Mean / EV	Mean	Median	
1	36	0.20	0.20	0.20	0.02	0.99	4	36	0.30	0.30	0.30	0.01	0.99	1.00	1.00	n/a
2	36	0.35	0.35	0.35	0.01	1.00	5	36	0.45	0.45	0.45	0.02	1.01	0.99	1.00	n/a
3	36	0.07	0.20	0.10	0.21	2.87	6	36	0.17	0.23	0.18	0.20	1.33	1.97	1.23	***
7	36	4.90	4.81	4.95	1.48	0.98	11	36	5.90	4.86	5.15	1.59	0.82	1.47	1.08	***
8	36	2.00	2.68	2.15	1.08	1.34	12	36	3.00	2.63	2.90	0.96	0.88	1.66	1.23	***
9	36	2.00	2.87	2.00	1.88	1.43	13	36	3.00	3.45	3.00	2.04	1.15	1.38	1.00	**
10	36	3.00	3.86	3.00	2.53	1.29	14	36	4.00	4.24	4.00	2.58	1.06	1.46	1.01	**

Table 2 *(continued)*

D) PZ experiment – B lotteries (Treatments 1, 2 and 3)

WTA valuation							WTP valuation							WTA /WTP[a]		Sign.[b]
Lott. No.	**Obs.**	**EV**	**Mean**	**Median**	**Std. Dev.**	**Mean / EV**	**Lott. No.**	**Obs.**	**EV**	**Mean**	**Median**	**Std. Dev.**	**Mean / EV**	**Mean**	**Median**	
1	38	0.20	0.20	0.20	0.00	1.00	4	38	0.30	0.30	0.30	0.00	1.00	1.00	1.00	n/a
2	38	0.35	0.35	0.35	0.01	1.00	5	38	0.45	0.45	0.45	0.01	0.99	1.00	1.00	n/a
3	38	0.43	0.44	0.45	0.17	1.01	6	38	0.53	0.49	0.50	0.18	0.92	1.13	1.07	**
7	38	2.10	2.67	2.10	1.56	1.27	11	38	3.10	2.41	2.48	0.76	0.78	1.67	1.36	***
8	38	3.00	2.80	3.00	0.99	0.93	12	38	4.00	3.10	3.00	1.07	0.78	1.34	1.20	***
9	38	2.00	2.69	2.00	1.81	1.34	13	38	3.00	2.67	3.00	1.24	0.89	1.97	1.34	***
10	38	7.00	6.78	7.00	1.70	0.97	14	38	8.00	7.03	7.41	2.11	0.88	1.20	1.08	**

E) PZ experiment – mugs

WTA valuation					WTP valuation					WTA /WTP[a]		Sign.[b]
	Obs.	**Mean**	**Median**	**Std. Dev.**	**Lott. No.**	**Obs.**	**Mean**	**Median**	**Std. Dev.**	**Mean**	**Median**	
Pooled	38	5.56	5.00	3.58	Pooled	36	6.62	6.00	4.20	0.84	0.83	
Mug last	24	5.48	5.00	3.40	Mug last	24	5.99	6.00	2.90	0.92	0.83	
Mug first	14	5.71	5.10	4.00	Mug first	12	7.88	6.50	6.00	0.72	0.78	

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 and KKT treatments (two-tail rank sum test).