

## How certain is the uncertainty effect?

Ondřej Rydval · Andreas Ortmann ·  
Sasha Prokosheva · Ralph Hertwig

Received: 27 December 2007 / Accepted: 22 September 2009 / Published online: 9 October 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** We replicate three tasks for which Gneezy, List and Wu (Q. J. Econ. 121(4):1283–1309, 2006) document the so-called uncertainty effect: People value a binary lottery over non-monetary outcomes less than other people value the lottery's worse outcome. While the authors implement verbal lottery descriptions, we use a physical lottery format and also provide subjects with complete information about the goods they are to value. We observe for all three pricing tasks that subjects' willingness to pay for the lottery is significantly higher than other subjects' willingness to pay for the lottery's worse outcome.

**Keywords** Decision making under risk · Framing · Experiments · Task ambiguity

**JEL Classification** C81 · C91 · C93 · D83

---

O. Rydval  
Max Planck Institute of Economics, Jena, Germany

O. Rydval · S. Prokosheva  
CERGE-EI, Prague, Czech Republic

A. Ortmann (✉)  
School of Economics, Australian School of Business, UNSW, Sydney, NSW 2052, Australia  
e-mail: [a.ortmann@unsw.edu.au](mailto:a.ortmann@unsw.edu.au)

A. Ortmann  
e-mail: [aortmann@yahoo.com](mailto:aortmann@yahoo.com)

S. Prokosheva  
Universidad Carlos III de Madrid, Madrid, Spain

R. Hertwig  
University of Basel, Basel, Switzerland

## 1 Introduction

Most theories of decision making under risk require that the value of any risky prospect lie between the value of the prospect's best and worst outcomes. Gneezy et al. (2006, henceforth GLW) term this requirement the *internality axiom* (henceforth IA) and document its systematic violations.<sup>1</sup> For various valuation goods (bookstore gift certificates, one-year deferred payments, sports cards, and work effort), elicitation modes (pricing and choice), and implementation variants (hypothetical and real-stakes, laboratory and field experiments), GLW demonstrate that people value equiprobable binary lotteries *less* than other people value the lotteries' worse outcomes.

GLW argue that the IA violations are due to an *uncertainty effect* triggered by two lottery design features that obstruct the IA. First, their lotteries involve non-monetary outcomes, the valuation of which may induce higher cognitive demands or perception of uncertainty. Second, the between-subjects design makes it less likely that subjects value the lotteries based on directly valuing their outcomes. GLW propose that subjects value the expectation of the outcomes' face values and discount the valuation for the risk involved in the lottery. This "risk and return" valuation process—incompatible with most theories of decision making under risk—could indeed explain the observed IA violations if high risk premia were levied on equiprobable or similar lotteries (as opposed to lotteries with more unequal probability mixes).<sup>2</sup>

We replicate three of GLW's tasks, namely hypothetical and real-stakes pricing of bookstore gift certificates, and hypothetical pricing of one-year deferred payments. We find no evidence for IA violations and, hence, for the *uncertainty effect*. In all three pricing tasks, subjects' willingness to pay for the lotteries is significantly higher than other subjects' willingness to pay for the lotteries' worse outcomes. It therefore seems that the *uncertainty effect* phenomenon is less robust than GLW's results suggest.

As discussed in Ortmann et al. (2007) and in Sect. 2.1 below, our replication study was initially motivated by the concern that GLW's experimental instructions—describing the lottery structure verbally and constructing a conceptual divide between the lottery ticket to be valued and the lottery outcomes—could have been misinterpreted by subjects, thus possibly contributing to the observed IA violations. In order to increase the transparency of the lottery structure, all but two of our lottery treatments use a "physical" format: Subjects state their willingness to pay for the opportunity of drawing a good (a gift certificate or a deferred payment form) from a closed bag containing two goods that are identical except for their face value. The experimenter physically demonstrates the lottery structure. Unlike GLW, we also equip subjects with complete information about the goods they are to value.

---

<sup>1</sup>In their footnote 1, GLW discuss theories of decision making under risk that in principle permit IA violations. While the IA is seemingly derived from deterministic theories, its empirical tests have implications for stochastic theories as well.

<sup>2</sup>In the hypothetical pricing of gift certificates, the IA is violated not only for equiprobable lotteries but also for lotteries with probability mixes of (0.6, 0.4) and (0.4, 0.6). For lotteries with more unequal probability mixes, the IA is not violated.

Several other studies have investigated IA violation in GLW's or closely related tasks. Keren and Willemsen (2009) replicate GLW's hypothetical pricing and choice of gift certificates, using various lottery descriptions and *ex-post* lottery comprehension tests. When the lottery is described as in GLW or is reworded but still purely verbal, the IA is violated or not supported, and the comprehension rates are only 29–43%. By contrast, when the lottery is described by a coin flip or a spinner wheel, the IA is overall weakly supported and the comprehension rates rise to 69–87%. In each task, IA violation seems to occur only for subsets of subjects failing the comprehension test. While these findings are illuminating, we note that GLW did use a coin-flip lottery description in their real-stakes gift-certificate task, yet observed IA violation. Also, Keren and Willemsen's coin-flip and spinner-wheel lottery descriptions are still verbal (compared to our physical lottery format) and retain the lottery-ticket or lottery-participation terminology, which might have contributed to the still noticeable proportion of subjects failing the comprehension test.

Simonsohn (2009, Experiment 1) conducts the hypothetical pricing of gift certificates, but his design differs from GLW in that the lotteries involve pairs of bookstore and restaurant gift certificates in various mixtures. Also, subjects in a baseline (certainty) treatment value both certificates involved in the corresponding lottery treatment.<sup>3</sup> Simonsohn further slightly rewords GLW's lottery instructions to make the lottery structure more transparent. Despite that, the IA is always violated, in most cases strongly. In Experiment 2, Simonsohn replicates GLW's lottery treatment for the hypothetical pricing of bookstore gift certificates, adapting the reworded lottery instructions from Experiment 1 and implementing two *ex-post* lottery comprehension tests. The comprehension rate turns out lower for subjects facing Keren and Willemsen's test (79%) compared to subjects facing a multiple-choice, logically equivalent test (92%).<sup>4</sup> The high comprehension rates in Experiment 2, combined with the extensive IA violation in the related Experiment 1, cast some doubt on the aforementioned conjecture that the IA violations observed in GLW were due to misinterpretation of the lottery instructions.<sup>5</sup>

Sonsino (2008) observes IA violations even in a *within*-subjects design. In his web-based experiment, subjects first value three gift certificates with widely different, *undisclosed* market prices,<sup>6</sup> having complete information about the certificates similar to our design. Then each subject values binary lotteries over pairs of the certificates while observing her previous valuations of the certificates themselves. The

---

<sup>3</sup>Based on evidence on valuing single versus multiple items, Simonsohn argues that if subjects in the baseline treatment observe both gift certificates, they will be likely to value the worse one lower, which could help alleviate IA violation.

<sup>4</sup>Simonsohn attributes this result to Keren and Willemsen's test being hard to understand. We note that the two comprehension tests are logically but perhaps not behaviorally equivalent, which might have also contributed to the difference in the comprehension rates.

<sup>5</sup>Simonsohn cannot assess whether IA violation occurred in Experiment 2 since he did not elicit the WTP for the worse gift certificate. The author draws further conclusions based on relating the subjects' WTP for the gift-certificate lottery to their WTP for one-year deferred payment (elicited before the lottery treatment). However, we see no *a priori* reason why the two WTP valuations should be related and thus find the author's conclusions unsubstantiated.

<sup>6</sup>The certificates feature a luxurious weekend vacation, a gourmand dinner, and a choice between a fine bottle of wine and a box of gourmet chocolate.

valuations are elicited using a sequence of six-bidder Vickrey auctions (the probability of an auction being played out is about 5%). The lotteries, involving various probability mixes, are described verbally using the lottery-ticket terminology and pie charts. Subjects are invited to participate in the subsequent lottery draw carried out by volunteer subjects secretly choosing numbers that are then compared to randomly generated numbers. Sonsino finds that in almost 12% of cases, subjects value a lottery less than they value either of the lottery's outcomes, and 27% of subjects do so at least once. IA violations occur more frequently when the probability of winning the lottery's better outcome is decreased, contrary to GLW's study where IA violations occur only for equiprobable or similar lotteries. Sonsino's striking results could at least partly stem from the web-based nature of the experiment (potentially compromising experimental control), the auction-based elicitation mechanism, or the sequential nature of the valuations possibly generating order effects. Also, the verbal lottery description and the lengthy description of the lottery draw could have been misinterpreted by some subjects and contributed to what Sonsino calls "lottery aversion."<sup>7</sup>

We see the main contribution of our replication study in the use of the physical lottery format. Combined with giving subjects complete information about the goods to be valued, our implementation reduces the scope for misinterpretation to a minimum and sets the ground for "clean" replication of GLW's tasks. We do not claim that the IA violations documented in GLW arose from misinterpretation, nor can we prove or disprove this conjecture. We are open to the possibility that the consistent differences between our, GLW's, and other findings arise from implementation differences detailed below.

Another virtue of our replication study is that we conduct it under relatively more controlled experimental conditions. Both Keren and Willemsen (2009) and Simonsohn (2009) conduct their replications as part of a series of experiments of unknown (related) nature and order, while Sonsino's (2008) experiment is web-based. Moreover, our most extensive replication task uses substantial monetary incentives, whereas the other replication studies were hypothetical or low-stakes, and hence likely afflicted by unnecessary noise (e.g., Hertwig and Ortmann 2001; Rydval and Ortmann 2004).

## 2 Design and implementation

We replicated three tasks for which GLW document IA violations: hypothetical pricing of bookstore gift certificates, real-stakes pricing of bookstore gift certificates, and hypothetical pricing of one-year deferred payments. For each task, a lottery treatment (L) elicited willingness to pay (henceforth WTP) for an equiprobable binary

---

<sup>7</sup>Sonsino (2008) argues that IA violation could be triggered by subjects' aversion to the presented lotteries *per se*. In his post-experimental questionnaire, "aversion to lotteries" was the most frequently chosen explanation for IA violation. Subjects were shown an example of within-subjects IA violation and, if admitting to the (hypothetical) possibility of exhibiting such behavior, were prompted to choose their preferred explanation for the behavior from a list of three options, the other two being "noise distraction" and "other explanations."

**Table 1** Design matrix of pricing tasks and treatments

Task	Lottery treatments	Baseline treatments
Hypothetical pricing of gift certificates	$L_{replication}$	
	$L_{rewording}$	
	$L_{physical}$	$B_{physical}$
Real-stakes pricing of gift certificates	$L_{sym}$	$B_{sym-wide}$
	$L_{asym}$	$B_{sym-narrow}$
Hypothetical pricing of deferred payments	$Ld_{physical}$	$Bd_{physical}$
	$Ld_{sym}$	$Bd_{sym-wide}$
	$Ld_{asym}$	$Bd_{sym-narrow}$

lottery over two gift certificates or deferred payments with a face value of  $x$  and  $2x$ . A corresponding baseline treatment (**B**) elicited WTP for the worse gift certificate or deferred payment with a face value of  $x$ .<sup>8</sup> As in GLW, all comparisons between a lottery treatment and a corresponding baseline treatment drew on a between-subjects design. Towards a better understanding of our results, we ran several lottery and baseline treatments for each task, as is explained below and outlined in Table 1.

## 2.1 Hypothetical pricing of bookstore gift certificates

As mentioned above, our study was initially motivated by the concern that some subjects might have misinterpreted GLW's lottery instructions. Hence we conducted a lottery treatment  $L_{replication}$  using GLW's instructions, where subjects valued a lottery ticket that had gift certificates worth  $x$  and  $2x$  as the equiprobable outcomes. In another lottery treatment  $L_{rewording}$ , we aimed to make the instructions less ambiguous by asking subjects to value a gift certificate worth either  $x$  or  $2x$  with equal probability (described as a fair-coin toss).<sup>9</sup> We used gift certificates for one of the largest bookstores in the Czech Republic (Luxor Book Palace) located within walking distance of the experimental site. As in GLW, the certificates were valid for the next two weeks. These initial lottery treatments, discussed in more detail in Ortman et al. (2007), appear at the top of Table 1.

We found that our rewording essentially eliminated the possibility of IA violation: 84% of valuations in  $L_{rewording}$  were at or above the face value of the worse gift certificate. Hence we thought that running a baseline treatment corresponding to  $L_{rewording}$  and  $L_{replication}$  was unnecessary. In retrospect, running this baseline treatment would have deepened our understanding of the conditions under which IA violation occurs, namely whether we could replicate GLW's results in  $L_{replication}$ . We address this issue in more detail in Sect. 3.1.

<sup>8</sup>Our design therefore does not address the possibility that lotteries are valued *more* than their best outcomes, though Sonsino (2008) shows this can happen in very rare cases (in 0.75% of observations).

<sup>9</sup>View the experimental instructions at [www.econ.mpg.de/english/staff/esi/rydval](http://www.econ.mpg.de/english/staff/esi/rydval).

As another potential caveat of our initial findings, we could not rule out the possibility that our rewording itself induced some form of misinterpretation of the lottery structure. In order to reduce the scope for misinterpretation to a minimum, we implemented  $L_{physical}$  and all additional lottery treatments in Table 1 using the following physical lottery format.<sup>10</sup>

While reading the lottery instructions aloud, the experimenter presented a bag into which he placed the two gift certificates and demonstrated how one certificate was to be randomly drawn from the closed bag. The instructions explained that the two certificates were identical except for their value (subjects were invited to verify the claim when inspecting the circulated certificates) and thus the chances of drawing either the better or the worse certificate would be equal. Subjects were then asked to state their WTP for the opportunity of drawing one gift certificate from the bag.

$L_{physical}$  and the corresponding baseline treatment  $B_{physical}$  also used Luxor Book Palace gift certificates which, however, were described in more detail than in  $L_{replication}$  and  $L_{rewording}$ . The instructions stated that the certificates were valid for the next three months and allowed in-store and online purchase of books (including CD and DVD formats), maps, stationery, etc.<sup>11</sup> The experimenter circulated among subjects several certificates with the appropriate face value in order to ensure common knowledge and enhance credibility. He also mentioned other conditions of use of the certificates, such as cash not being returned if a purchase falls below the certificates' face value, and the possibility of using the certificates to co-finance purchases exceeding their face value.

Following GLW, our instructions asked subjects to state the highest price they would be willing to pay. One can imagine a procedurally more incentive-compatible mechanism for hypothetical WTP elicitation, such as a hypothetical-stakes version of the real-stakes mechanism described in the next section. However, we wanted to retain the key features of GLW's hypothetical-stakes design for which the authors observed IA violations. We hoped—as implicitly did the authors of the other aforementioned hypothetical studies—that the (potential) hypothetical bias would not interact with the lottery and baseline treatments in a way that biases the treatment effect towards or away from IA violation.

## 2.2 Real-stakes pricing of bookstore gift certificates

Except for the real-stakes implementation, the lottery and baseline treatments in the penultimate row of Table 1 were constructed in analogy to  $L_{physical}$  and  $B_{physical}$ , respectively. That is, the lottery treatments were implemented using the physical lottery

<sup>10</sup>Using a physical lottery format to study sources of risky decision anomalies is not a novel idea. In Grether and Plott (1979, Sect. 2), for instance, a physical lottery demonstration did not alleviate preference reversals. In Gigerenzer et al. (1988, Experiment 1), by contrast, a physical demonstration of random sampling lessened base rate neglect significantly.

<sup>11</sup>The three-month validity period of the gift certificates is longer than the two-week period in our initial two lottery treatments, GLW, Keren and Willemsen (2009), and presumably also Simonsohn (2009). The discrepancy was imposed on us by the bookstore's current policy. For completeness, Sonsino's (2008) gift certificates were valid for the next 6 months.

format,<sup>12</sup> and all treatments (lottery and baseline) provided subjects with complete information about the gift certificates (see the previous section for details).

We used the multiple price list (MPL) mechanism to elicit WTP. In treatments  $\mathbf{L}_{sym}$  and  $\mathbf{B}_{sym-wide}$ , subjects received  $2x$  for participating. They were asked to indicate (by circling either Yes or No) their willingness to pay the listed prices rising from  $0.1x$  to  $2x$  in  $0.1x$  increments. Subjects were informed that only one of their 20 decisions would be payoff-relevant. Each subject eventually randomly drew a card from a box with cards numbered 1 to 20 to determine the payoff-relevant row. If the subject circled Yes in that row, she paid the price and received the gift certificate worth  $x$  (in  $\mathbf{B}_{sym-wide}$ ) or randomly drew a certificate from a bag containing two certificates worth  $x$  and  $2x$  (in  $\mathbf{L}_{sym}$ ). If she circled No in the payoff-relevant row, she earned the participation fee.

We ran a supplementary baseline treatment  $\mathbf{B}_{sym-narrow}$  with a narrower MPL ranging from  $0.1x$  to  $x$ ; subjects received  $x$  for participating and made only 10 Yes-or-No decisions. This MPL procedure seems more natural for valuing a good worth  $x$ . However, if subjects were drawn towards the middle of the MPL,  $\mathbf{B}_{sym-narrow}$  could (*ceteris paribus*) induce lower WTP valuations and hence work against the occurrence of IA violation. We accounted for the potential bias caused by the “mid-table” effect by comparing valuations in  $\mathbf{L}_{sym}$  and  $\mathbf{B}_{sym-wide}$  which both have the same wide MPL. Moreover, comparing valuations in  $\mathbf{B}_{sym-wide}$  and  $\mathbf{B}_{sym-narrow}$  told us whether the mid-table effect was actually present in the baseline treatments.

We also conducted a supplementary lottery treatment  $\mathbf{L}_{asym}$  where, similar to Andersen et al. (2007) and Harrison et al. (2007), the MPL was asymmetrically “skewed low” in that the mid-row price was below the mean of the MPL (in fact below  $0.8x$ ). If the mid-table effect were present,  $\mathbf{L}_{asym}$  should (*ceteris paribus*) induce lower WTP valuations and hence favor IA violation. Thus one reason for implementing  $\mathbf{L}_{asym}$  was to give IA violation a fair chance to occur in our data. Furthermore, comparing valuations in  $\mathbf{L}_{sym}$  and  $\mathbf{L}_{asym}$  told us whether the mid-table effect was present in the lottery treatments.<sup>13,14</sup>

We preferred the MPL mechanism to the Becker-DeGroot-Marschak (BDM) mechanism used by GLW. Any incentive-compatible mechanism should serve the goal of assessing the direction of the lottery-baseline treatment effect (i.e., the occurrence of IA violation or lack thereof), *unless* the mechanism interacts with the lottery and baseline treatments in a way that biases the treatment effect. Given the well-known concerns with the BDM mechanism (e.g., Karni and Safra 1987; Harrison 1992; Horowitz 2006), and having explicitly accounted for the potential

<sup>12</sup>We were unable to obtain gift certificates worth  $2x$ , so we used two identical table tennis balls marked  $x$  and  $2x$  to represent the real gift certificates. Subjects were informed that if they drew a ball marked  $2x$ , they would receive two gift certificates worth  $x$ .

<sup>13</sup>Up to  $1.4x$ , the MPL in  $\mathbf{L}_{asym}$  and  $\mathbf{L}_{sym}$  share all prices, and even above  $1.4x$ , it turns out that no subject in  $\mathbf{L}_{sym}$  made a valuation that could not have been made in  $\mathbf{L}_{asym}$ . In this sense, comparing the two WTP distributions is innocuous.

<sup>14</sup>While our variation of the real-stakes lottery and baseline treatments may have induced other behavioral effects, such as changing the effective power of financial incentives, we cannot find a reason why any such effect should work against the occurrence of IA violation.

bias caused by the mid-table effect, we concluded that the MPL mechanism was a less problematic choice.

There are refinements of our basic MPL procedure that, for example, allow subjects to express indifference and elicit WTP more precisely by iteratively decreasing the MPL price increments (e.g., Andersen et al. 2007). However, given our focus on the direction of the lottery-baseline treatment effect (rather than on WTP point estimates), the benefits of such refinements in our view do not outweigh their potential costs arising from the increased complexity of the elicitation procedure. Our MPL price increments of 0.1x (about \$1) seem fine enough to lessen concerns related to the interval-censored nature of MPL responses, which we in any case address statistically.<sup>15</sup> Also, our subjects could fill out the MPL in a non-monotonic manner to (among other things) indicate indifference. We tackle these (very rare) cases by using wider price intervals to represent the affected subjects' WTP, and we always do so in a way that favors the occurrence of IA violation.

### 2.3 Hypothetical pricing of one-year deferred payments

Except for different goods to be valued, the lottery and baseline treatments in the last row of Table 1 were constructed in analogy to  $L_{physical}$  and  $B_{physical}$ , respectively: All the lottery treatments were implemented using the physical lottery format, and we asked subjects in all treatments (lottery and baseline) to state the highest price they would be willing to pay (see Sect. 2.1 for details).

We used hypothetical payment forms guaranteeing cash payment in one year from the date of the experiment. The instructions explained that the deferred payment would (in a real-stakes scenario) be guaranteed by the research organization financing the experiment. We circulated several (hypothetically filled out) payment forms of the research organization for subjects' inspection. While reading the instructions aloud, the experimenter also mentioned that the deferred cash payment would (in a real-stakes scenario) be made by one of the experimenters at the experimental site.

The deferred-payment task contained a within-subjects component: It was conducted as a "surprise" task directly following the hypothetical and real-stakes pricing of gift certificates. In particular, the deferred-payment lottery treatments  $Ld_{physical}$ ,  $Ld_{sym}$  and  $Ld_{asym}$  differed from each other only in that they followed different gift-certificate lottery treatments, namely  $L_{physical}$ ,  $L_{sym}$  and  $L_{asym}$ , respectively. Similarly, the deferred-payment baseline treatments  $Bd_{physical}$ ,  $Bd_{sym-wide}$  and  $Bd_{sym-narrow}$  were conducted after subjects completed  $B_{physical}$ ,  $B_{sym-wide}$  and  $B_{sym-narrow}$ , respectively.

While the instructions for the deferred-payment treatments reminded subjects that they faced a new task unrelated to the gift-certificate treatments just completed, the gift-certificate valuation undoubtedly influenced the subsequent deferred-payment valuation in some manner. Given the variety of the gift-certificate lottery (baseline) treatments preceding the deferred-payment lottery (baseline) treatments, we can

<sup>15</sup>There is also controversy as to whether WTP can actually be elicited as precisely as required by mechanisms eliciting point-estimate responses, such as the BDM mechanism or the hypothetical elicitation mechanism used in our and GLW's hypothetical pricing tasks. See, for example, Andersen et al. (2007) for further discussion.

study this influence in a systematic manner in Sect. 3.3. For this reason, we do not view the deferred-payment valuations as less informative than the gift-certificate valuations.

## 2.4 Other implementation details

The experimental sessions were run in a pen-and-paper format (as in GLW) and were conducted in Czech by the first author. All parts of the experiment were anonymous and payments as well as lottery draws (if any) were done privately outside the lab at the end of a session. Including an initial demographic questionnaire, hypothetical sessions lasted about 20 minutes while real-stakes sessions lasted slightly longer. Subjects earned 100 CZK (about \$5) for participating in the hypothetical sessions; the participation fee was 400 CZK or 200 CZK in sessions involving the real-stakes task (see Sect. 2.2 for reasons).

In the hypothetical pricing of gift certificates,  $x = 500$  CZK (about \$25), thus roughly matching the (student) purchasing power of  $x = \$50$  used in GLW. In the hypothetical pricing of deferred payments, we also used  $x = 500$  CZK to make the two hypothetical pricing tasks comparable (GLW used  $x = \$100$ ). In the real-stakes pricing of gift certificates,  $x = 200$  CZK (about \$10) and all subjects' decisions were played out, while GLW used  $x = \$50$  and played out the decisions of 5% of the subjects (one subject per session). Note that  $x = 200$  CZK is still a substantial amount of money, exceeding the price of almost half of the books and textbooks and a much larger proportion of the stationery items sold at the bookstore's online shop.

The initial two lottery treatments ( $L_{\text{replication}}$  and  $L_{\text{rewording}}$ ) were conducted in early April 2007. The subjects were 64 students from the Faculty of Social Sciences of the Charles University in Prague, recruited using posters. The other treatments presented in this study were conducted in early December 2008.<sup>16</sup> The subjects were 150 students from various Prague universities recruited online using ORSEE (Greiner 2004). Just above a third of them were students from various branches of engineering, one quarter were students of finance, business, management or accounting, another quarter were economists, and the remainder came from other fields. Overall, subjects were 18 to 30 years old with the mean and median age of 22 years, and 67% of them were males. No subject participated in our replication study more than once.

## 3 Results

Table 2 displays WTP summary statistics for the various treatments. To facilitate casual comparison of valuations across tasks, all WTP figures are reported as percentages of  $x$ , the face value of the worse gift certificate or deferred payment. Table 3 reports unconditional tests for the lottery-baseline treatment effect ( $t$ -test, Wilcoxon rank-sum test and Kolmogorov-Smirnov test) as well as Wald tests that condition on

---

<sup>16</sup>There were nine sessions, conducted in a quasi-random (but predetermined) order over four consecutive working days.

**Table 2** Summary statistics for willingness to pay<sup>a</sup>

Treatment	# of subjects	Mean WTP (std. dev.)	Median WTP	95% C.I. for WTP mean	95% binomial exact C.I. for WTP median
<b>L<sub>replication</sub></b>	32	90.31 (32.89)	100.00	(78.46, 102.17)	(60.00, 100.00)
<b>L<sub>rewording</sub></b>	32	111.03 (25.87)	105.00	(101.70, 120.35)	(100.00, 130.00)
<b>L<sub>physical</sub></b>	26	88.22 (34.00)	98.90	(74.49, 101.95)	(80.00, 100.00)
<b>B<sub>physical</sub></b>	15	69.20 (18.09)	70.00	(59.18, 79.22)	(60.00, 84.00)
<b>L<sub>sym</sub></b>	35	86.57 (36.21)	100.00	(74.13, 99.01)	(80.00, 100.00)
<b>L<sub>asym</sub></b>	18	78.61 (24.96)	90.00	(66.20, 91.02)	(60.00, 100.00)
<b>B<sub>sym-wide</sub></b>	39	66.67 (23.32)	70.00	(59.11, 74.23)	(50.00, 80.00)
<b>B<sub>sym-narrow</sub></b>	14	53.57 (24.37)	55.00	(39.50, 67.64)	(30.00, 80.00)
<b>L<sub>sym</sub> &amp; L<sub>asym</sub></b>	53	83.87 (32.80)	100.00	(74.83, 92.91)	(75.00, 100.00)
<b>B<sub>sym-wide</sub> &amp; B<sub>sym-narrow</sub></b>	53	63.21 (24.08)	60.00	(56.57, 69.85)	(50.00, 70.00)
<b>L<sub>d,physical</sub></b>	26	85.87 (39.10)	85.00	(70.07, 101.66)	(60.00, 120.00)
<b>L<sub>d,sym</sub></b>	36	77.42 (41.83)	85.00	(63.26, 91.57)	(60.00, 100.00)
<b>L<sub>d,asym</sub></b>	18	72.11 (33.23)	80.00	(55.59, 88.64)	(60.00, 94.00)
<b>B<sub>d,physical</sub></b>	15	61.27 (25.60)	70.00	(47.09, 75.44)	(50.00, 80.00)
<b>B<sub>d,sym-wide</sub></b>	39	45.81 (28.47)	50.00	(36.59, 55.04)	(36.00, 60.00)
<b>B<sub>d,sym-narrow</sub></b>	16	42.00 (25.91)	40.00	(28.19, 55.81)	(20.00, 64.00)
<b>L<sub>d,physical</sub> &amp; L<sub>d,sym</sub> &amp; L<sub>d,asym</sub></b>	80	78.97 (39.04)	80.00	(70.28, 87.66)	(60.00, 99.00)
<b>B<sub>d,physical</sub> &amp; B<sub>d,sym-wide</sub> &amp; B<sub>d,sym-narrow</sub></b>	70	48.25 (27.83)	50.00	(41.62, 54.89)	(40.00, 60.00)

<sup>a</sup>WTP figures are percentages of *x*, the face value of the worse gift certificate or deferred payment

**Table 3** Tests of the lottery-baseline treatment effect for willingness to pay<sup>b</sup>

Row	Null hypothesis	<i>t</i> -test without controls	Wald test with controls	Wilcoxon rank-sum test	Kolmogorov-Smirnov test
1	$L_{\text{rewording}} - L_{\text{replication}} = 0$	2.80 (0.0068)	2.93 (0.004)	2.56 (0.0105)	0.34 (0.045)
2	$L_{\text{physical}} - B_{\text{physical}} = 0$	2.34 (0.0247)	2.47 (0.015)	2.25 (0.0243)	0.47 (0.018)
3	$L_{\text{sym}} - L_{\text{asym}} = 0$	0.83 (0.4080)	0.98 (0.329)	1.40 (0.1625)	0.23 (0.485)
4	$B_{\text{sym-wide}} - B_{\text{sym-narrow}} = 0$	1.78 (0.0808)	1.49 (0.137)	1.67 (0.0950)	0.28 (0.322)
5	$(L_{\text{sym}} \& L_{\text{asym}}) - (B_{\text{sym-wide}} \& B_{\text{sym-narrow}}) = 0$	3.70 (0.0004)	4.06 (0.000)	3.90 (0.0001)	0.45 (0.000)
6	$(L_{\text{sym}} \& L_{\text{asym}}) - B_{\text{sym-wide}} = 0$	2.94 (0.0042)	3.45 (0.001)	3.18 (0.0015)	0.43 (0.001)
7	$L_{\text{asym}} - B_{\text{sym-wide}} = 0$	1.76 (0.0843)	1.99 (0.047)	1.87 (0.0613)	0.30 (0.165)
8	$(L_{\text{dphysical}} \& L_{\text{dsym}} \& L_{\text{dasym}}) - (B_{\text{dphysical}} \& B_{\text{dsym-wide}} \& B_{\text{dsym-narrow}}) = 0$	5.60 (0.0000)	6.65 (0.000)	4.97 (0.0000)	0.43 (0.000)
9	$(L_{\text{dsym}} \& L_{\text{dasym}}) - (B_{\text{dsym-wide}} \& B_{\text{dsym-narrow}}) = 0$	4.78 (0.0000)	5.37 (0.000)	4.29 (0.0000)	0.43 (0.000)
10	$L_{\text{dphysical}} - B_{\text{dphysical}} = 0$	2.18 (0.0357)	3.17 (0.002)	1.95 (0.0515)	0.43 (0.037)
11	$(L_{\text{dphysical}} \& B_{\text{dphysical}}) - (L_{\text{dsym}} \& L_{\text{dasym}} \& B_{\text{dsym-wide}} \& B_{\text{dsym-narrow}}) = 0$	2.50 (0.0137)	2.30 (0.023)	2.35 (0.0190)	0.21 (0.108)

<sup>b</sup>The “&” sign denotes pooling treatments. Two-sided test statistics (rounded to 2 d.p.) are followed by *p*-values in parentheses (precision as reported by *Stata*). Exact *p*-values are reported for the Kolmogorov-Smirnov test. The Wald test statistics are from OLS estimation (Row 1–2 & 8–11) and interval regression estimation (Row 3–7) with heteroskedasticity-robust standard errors. The collected demographic characteristics are jointly significant at the 5% level (Row 1–2) and 10% level (Rows 3–11). Subjects’ year of study (Row 1–2), age (Row 3–7) and field of study (Row 8–11) are significant at the 5% level; other controls including session dummies are individually insignificant at the 10% level. We omit the wealth proxy from the final estimations since it could be viewed as a controversial wealth indicator and is always highly insignificant

the collected demographic characteristics (age, gender, year and field of study, and a wealth proxy based on family car ownership).<sup>17</sup>

### 3.1 Hypothetical pricing of gift certificates

As mentioned above, we initially conducted the lottery treatments  $L_{replication}$  and  $L_{rewording}$ . Their implementation was closest to that of GLW, but differed from all subsequent treatments in terms of lottery representation, the gift certificates' description and validity period, subject pool composition, and implementation date (see Sects. 2.1 and 2.4). Since these implementation differences could affect valuations of gift certificates, we report the results for these initial lottery treatments separately.

WTP is significantly higher in  $L_{rewording}$  compared to  $L_{replication}$ , as shown in Row 1 of Table 3. Moreover, Table 2 displays that the confidence intervals for WTP mean and median in  $L_{rewording}$  are above the face value of the worse gift certificate. Inspection of the individual-level data further reveals that 84% of the WTP valuations match or exceed the face value. We originally viewed these results as demonstrating sufficiently that  $L_{rewording}$  eliminated the possibility of IA violation, and did not run the baseline treatment corresponding to  $L_{rewording}$  and  $L_{replication}$ . As a consequence, we cannot directly assess whether the IA is violated for  $L_{replication}$ . Yet, since the WTP valuations in  $L_{replication}$  are centered just below the face value of the worse gift certificate and 53% of them match or exceed the face value, observing IA violation for  $L_{replication}$  would have been possible but unlikely.

As mentioned in Sect. 2.1, we cannot completely rule out the possibility that the high valuations in  $L_{rewording}$  resulted from misinterpretation of our reworded lottery instructions. Thus we now turn to evidence from treatments using the physical lottery format. As shown in Row 2 of Table 3, WTP is significantly higher in  $L_{physical}$  than in  $B_{physical}$ . While clearly supporting the IA, one should not overstate the result given the relatively small sample sizes involved, especially in  $B_{physical}$ . The main purpose of conducting  $L_{physical}$  and  $B_{physical}$  was to complement the suggestive yet debatable evidence from  $L_{replication}$  and  $L_{rewording}$ . In our mind, the strongest evidence in support of the IA comes from the real-stakes gift-certificate task reported on in the next section.

### 3.2 Real-stakes pricing of gift certificates<sup>18</sup>

We first assess the extent of the mid-table effect.  $L_{asym}$  has lower WTP valuations compared to  $L_{sym}$ , which points in the direction of the mid-table effect. However, the difference is small and far from significant, as shown in Row 3 of Table 3. There is more sizeable evidence for the mid-table effect in the baseline treatments, but WTP

<sup>17</sup>Some of the tests may be deemed more appropriate than others depending on how one views the nature of the data—see also footnote 15.

<sup>18</sup>We excluded three incomplete responses that clearly indicated misunderstanding of the MPL valuation procedure, one in a lottery treatment and two in a baseline treatment. Also, two subjects entered a non-monotonic MPL response, namely a single No response surrounded by Yes responses. For these one-off cases of non-monotonicity, we recoded the non-monotonic No response as a Yes response, which favors IA violation since both subjects happened to be in a baseline treatment.

valuations are not significantly lower in  $\mathbf{B}_{sym-narrow}$  than in  $\mathbf{B}_{sym-wide}$ , as shown in Row 4 of Table 3.

The lottery-baseline treatment effect can be assessed in several ways. One can pool lottery treatments  $\mathbf{L}_{sym}$  and  $\mathbf{L}_{asym}$  and baseline treatments  $\mathbf{B}_{sym-wide}$  and  $\mathbf{B}_{sym-narrow}$ , which yields confidence intervals for WTP mean of (74.83, 92.91) and (56.57, 69.85), respectively (see Table 2). Hence the pooled lottery-baseline treatment effect is clearly in the direction of the IA, as confirmed in Row 5 of Table 3. Alternatively, one can make stricter treatment comparisons that favor IA violation, for example by excluding  $\mathbf{B}_{sym-narrow}$  (see Row 6 of Table 3), or by contrasting only  $\mathbf{L}_{asym}$  with  $\mathbf{B}_{sym-wide}$  (see Row 7 of Table 3). Even these stricter comparisons provide support for the IA.

A casual comparison of the hypothetical and real-stakes gift-certificate valuations (excluding  $\mathbf{L}_{replication}$  and  $\mathbf{L}_{rewording}$  from the comparison for reasons outlined in the previous section) reveals that there is a minor upward hypothetical bias in both the lottery and baseline treatments, in line with the direction in which hypothetical bias typically seems to work (e.g., Harrison and Rutstroem 2008). By contrast, a similar comparison of GLW's hypothetical and real-stakes gift-certificate valuations suggests a much stronger downward hypothetical bias in both the lottery and baseline treatments.

### 3.3 Hypothetical pricing of deferred payments

The lottery-baseline treatment effect can again be assessed in several ways. Following up on the lottery-baseline comparisons performed in the previous two sections, Table 2 shows that WTP valuations are higher in  $\mathbf{Ld}_{physical}$  compared to  $\mathbf{Bd}_{physical}$ , and also in both  $\mathbf{Ld}_{sym}$  and  $\mathbf{Ld}_{asym}$  compared to  $\mathbf{Bd}_{sym-wide}$  and  $\mathbf{Bd}_{sym-narrow}$ . Alternatively, pooling all lottery treatments and all baseline treatments yields confidence intervals for WTP mean of (70.28, 87.66) and (41.62, 54.89), respectively, as shown in the last two rows of Table 2. Hence the pooled lottery-baseline treatment effect is clearly in the direction of the IA, as confirmed in Row 8 of Table 3. One can also make a separate lottery-baseline comparison for the deferred-payment treatments following the real-stakes gift-certificate task (see Row 9 of Table 3) and those following the hypothetical gift-certificate task (see Row 10 of Table 3), which also provides support for the IA.

We next examine the potential influence of subjects' gift-certificate valuations on their subsequent deferred-payment valuations. Casual inspection of Table 2 suggests that the mean WTP is always lower in a deferred-payment treatment compared to the preceding gift-certificate treatment. Further, in terms of the mean WTP, the deferred-payment treatments are ordered identically to the gift-certificate treatments. As another aggregate-level observation, deferred-payment valuations are overall higher when preceded by hypothetical rather than real-stakes gift-certificate valuations (see Row 11 of Table 3), which could be a corollary of the minor upward hypothetical bias discussed in Sect. 3.2. At the individual level, however, subjects' deferred-payment valuations do not simply reflect their preceding gift-certificate valuations: Spearman's correlation coefficient between the two valuations ranges (across treatments) between 0.27 and 0.60. The correlation is generally higher for lottery treatments than for the

corresponding baseline treatments, which may be due to risk attitudes affecting both of a subject's lottery valuations whereas the baseline valuations lack this common factor.

#### 4 Discussion and conclusion

We observe that the internality axiom is supported in all three tasks: Subjects' willingness to pay for equiprobable binary lotteries is significantly higher than other subjects' willingness to pay for the lotteries' worse outcomes, regardless of whether subjects' valuations are real-stakes or hypothetical and whether the outcomes are gift certificates or deferred payments.

Our replication of GLW's tasks rests on using a physical lottery format and on providing subjects with complete information about the goods they are to value, which arguably renders misinterpretation of the tasks unlikely. However, we do not (and cannot) conclude that these implementation differences, rather than other ones, are primarily responsible for why our results differ from those of GLW and the other aforementioned replication studies.<sup>19</sup>

For instance, we had no choice but using different local bookstore gift certificates with a longer validity period, and different language of instructions. The calendar time of implementing the task could also play a role, though our results point in the same direction for the two different implementation dates. Our findings could also be specific to our subject pool, though conditioning on the collected demographic characteristics leaves the results qualitatively unchanged. Naturally, replicating our implementation of GLW's tasks in other subject pools would provide a useful robustness check of our results.

We do not want to draw too strong a conclusion about the reality of internality axiom violation and hence the *uncertainty effect*. Nor do we wish to dispute GLW's view that their tasks involve high cognitive demands and perception of uncertainty, possibly triggering a lottery valuation process incompatible with normative risky decision theories, or triggering what Sonsino (2008) calls lottery aversion. Nevertheless, in our implementation of GLW's tasks, lottery valuation processes incompatible with the internality axiom seem not prevalent.

Converging lines of research in psychology suggest that "experiencing" information (e.g., sampling outcomes from payoff distributions, experiencing base rates, etc.), as opposed to encountering symbolic descriptions thereof, can strongly influence people's reasoning and judgments (e.g., Gigerenzer et al. 1988; Weber et al. 1993; Koehler 1996; Hertwig et al. 2004). The effect of our physical lottery implementation could be another manifestation of this finding.

**Acknowledgements** We thank Pavlo Blavatsky, Jordi Brandts, John Duffy, Uri Gneezy, Glenn Harrison, John List, Doron Sonsino, Tim Salmon, Lise Vesterlund, George Wu, participants of the ESA World Meeting 2007 at LUISS in Rome, and an anonymous referee for comments. This research was partly

<sup>19</sup>IA violation could in principle arise from genuinely higher gift-certificate or deferred-payment valuations in a baseline treatment compared to a corresponding lottery treatment, especially if coupled with high risk aversion of subjects in the lottery treatment. However, this scenario cannot account for the systematic differences between our results and those of other studies.

supported by a research center grant No. LC542 of the Ministry of Education of the Czech Republic implemented at CERGE-EI—a joint workplace of the Center for Economic Research and Graduate Education, Charles University, Prague, and the Economics Institute of the Academy of Sciences of the Czech Republic. Ortmann (Senior Researcher at the Economics Institute) acknowledges support of the grant No. MSM0021620846 from the Czech Ministry of Education, Youth, and Sports. Hertwig acknowledges support of the Swiss National Science Foundation Grant 100014-118283. The usual caveats apply.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Andersen, S., Harrison, G. W., Lau, M. I., & Rutstrom, E. E. (2007). Valuation using multiple price list formats. *Applied Economics*, *39*(6), 675–682.
- Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 513–525.
- Gneezy, U., List, J. A., & Wu, G. (2006). The uncertainty effect: When a risky prospect is valued less than its worst outcome. *Quarterly Journal of Economics*, *121*(4), 1283–1309.
- Greiner, B. (2004). *The online recruitment system ORSEE 2.0—A guide for the organization of experiments in economics*. University of Cologne, Working Paper Series in Economics 10.
- Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, *69*(4), 623–638.
- Harrison, G. W. (1992). Theory and misbehavior of first-price auctions: Reply. *American Economic Review*, *82*(5), 1426–1443.
- Harrison, G. W., & Rutstrom, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results* (Vol. 1). Amsterdam: North-Holland.
- Harrison, G. W., List, J. A., & Towe, C. (2007). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. *Econometrica*, *75*(2), 433–458.
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–402.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.
- Horowitz, J. K. (2006). The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, *93*(1), 6–11.
- Karni, E., & Safra, Z. (1987). Preference reversals and the observability of preferences by experimental methods. *Econometrica*, *55*(3), 675–685.
- Keren, G., & Willemsen, M. C. (2009). Decision anomalies, experimenter assumptions, and participants' comprehension: Re-evaluating the uncertainty effect. *Journal of Behavioral Decision Making*, *22*(3), 301–317.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–53.
- Ortmann, A., Prokosheva, S., Rydval, O., & Hertwig, R. (2007). *Valuing a risky prospect less than its worst outcome: Uncertainty effect or task ambiguity?* Jena Economic Research Paper 2007-038 and CERGE-EI Working Paper 334.
- Rydval, O., & Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: An illustration. *Economics Letters*, *85*(3), 315–320.
- Simonsohn, U. (2009). Direct risk aversion: Evidence from risky prospects valued below their worst outcome. *Psychological Science*, *20*(6), 686–692.
- Sonsino, D. (2008). Disappointment aversion in internet bidding-decisions. *Theory and Decision*, *64*(2–3), 363–393.
- Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1151–1164.