This article was downloaded by: [155.246.103.35] On: 05 April 2017, At: 17:58
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Do Preference Reversals Disappear When We Allow for Probabilistic Choice?

Graham Loomes, Ganna Pogrebna

To cite this article:
Graham Loomes, Ganna Pogrebna (2017) Do Preference Reversals Disappear When We Allow for Probabilistic Choice?. Management Science 63(1):166-184. http://dx.doi.org/10.1287/mnsc.2015.2333

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2017, The Author(s)

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 1, January 2017, pp. 166–184
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Do Preference Reversals Disappear When We Allow for Probabilistic Choice?

Graham Loomes,[a] **Ganna Pogrebna**[b]

[a] Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom; [b] International Institute for Product and Service Innovation, Warwick Manufacturing Group, University of Warwick, Coventry CV4 7AL, United Kingdom

Contact: g.loomes@warwick.ac.uk (GL); g.pogrebna@warwick.ac.uk (GP)

Received: February 13, 2015
Accepted: August 14, 2015
Published Online in Articles in Advance: March 14, 2016

https://doi.org/10.1287/mnsc.2015.2333

Copyright: © 2017 The Author(s)

Abstract. The "preference reversal phenomenon," a systematic disparity between people's valuations and choices, poses challenges for theory and policy. Using a very general formulation of probabilistic preferences, we show that the phenomenon is not mainly due to intransitive choice. We find a high degree of regularity *within* choice tasks and also *within* valuation tasks, but the two types of tasks appear to evoke very different cognitive processes, even when the experimental environment tries to minimise differences. We discuss possible implications for modelling and eliciting preferences.

History: Accepted by Manel Baucells, decision analysis.

Open Access Statement: This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "*Management Science*. Copyright © 2017 The Author(s). https://doi.org/10.1287/mnsc.2015.2333, used under a Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/."

Funding: G. Loomes acknowledges support from the UK Economic and Social Research Council [Grants RES-051-27-0248 and ES/K002201/1] and from the Leverhulme Trust "Value" Programme [RP2012-V-022]. G. Pogrebna acknowledges financial support from the Leverhulme Trust under the Early Career Fellowship scheme and under Research Councils UK/Engineering and Physical Sciences Research Council [Grant EPL023911/1].

Supplemental Material: Data are available at https://doi.org/10.1287/mnsc.2015.2333.

Keywords: preference reversal phenomenon • probabilistic preferences • stochastic choice

1. Introduction

More than three decades ago, Grether and Plott (1979) drew economists' attention to an unsettling regularity reported by experimental psychologists (Lichtenstein and Slovic 1971, Lindman 1971). The regularity in question—the *preference reversal phenomenon*—took the following form. Two lotteries were constructed: a "$-bet," which offered a relatively high payoff with a probability well below 0.5, and a "*P*-bet," which offered a considerably higher probability of a more modest payoff. Participants in the experiment were asked to do three things: to place a certainty equivalent value[1] upon the $-bet, to place a certainty equivalent value upon the *P*-bet, and to make a straight choice between the two. Most conventional decision theories suppose that an individual who prefers one bet to the other will pick the preferred option when offered a choice between the two and will also place a higher certainty equivalent value on whichever option he or she prefers. However, Lichtenstein and Slovic (1971) reported that a substantial proportion of their experimental participants flouted this expectation by choosing the *P*-bet while placing a strictly higher value on the $-bet. The opposite "anomaly"—choosing the $-bet while placing a higher value on the *P*-bet—was rarely observed. It is this asymmetry that constitutes the classic preference reversal (PR) pattern.

Grether and Plott (1979) had initially supposed that this phenomenon would disappear—or at least, be greatly attenuated—if stronger incentives and stricter experimental controls were deployed. But, in fact, the phenomenon persisted, and many other experiments since that time have found this classic PR pattern of behaviour to be easy to replicate and quite difficult to eliminate without considerable effort and/or supplementary mechanisms; see the survey by Seidl (2002).

At the level of modelling, this phenomenon would appear to present a fundamental challenge to general theories that assume transitivity. At the level of practical application, it raises concerns about the use of stated values as a basis for guiding public policy: if such patterns also occur in the domain of, say, environmental goods, there is a danger that using stated willingness to pay in cost–benefit analysis might lead to priority being given to projects that would not be chosen if citizens were in a position to express choices or rankings directly. For reasons of both theory and policy, therefore, it is important to gain a better understanding of what this phenomenon really represents in terms of the structure of human preferences

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

and/or the validity of different procedures for eliciting those preferences.

This paper aims to contribute to a better understanding of those issues. The next section discusses in more detail the main competing explanations of PR and some of the evidence for and against the different accounts. That discussion raises issues about the noisiness or imprecision of many people's responses, so in Section 3 we outline a (deliberately broad) framework of probabilistic preferences within which our inquiry will be conducted. In Section 4 we report two substantial experiments. The data generated by these experiments strongly suggest that for pairs of bets typical of the PR literature, most people's preferences exhibit some degree of stochastic variability, but their "core" preferences mostly conform with the probabilistic formulation of transitivity known as weak stochastic transitivity. Put another way, when certainty equivalent values are inferred from repeated binary choices, the classic PR phenomenon largely disappears, and the reversals that remain are relatively few in number and small in magnitude. By contrast, when certainty equivalent values are elicited more directly via a standard incentive-compatible mechanism, the same individuals display a very strong PR pattern of the usual kind. Our results broadly support the conclusions drawn by Tversky et al. (1990) and Bostic et al. (1990), although we find "overvaluation" of all bets rather than the mix of overvaluation and undervaluation that those studies report.

It is not possible to reconcile the data with conventional deterministic models—even those that permit some level of preference reversal (e.g., Loomes and Sugden 1983). So in Section 5 we explore the data we have generated to see whether we can shed further light on the disparities between choice and valuation, and we discuss whether we can account for them in terms of a model of probabilistic choice known as decision field theory (Busemeyer and Townsend 1993). It turns out that this model—at least, as applied by Johnson and Busemeyer (2005)—cannot adequately accommodate our data, so in the final section we discuss some possible directions for future theoretical development and some implications for applied research.

2. Different Possible Explanations of the Classic PR Pattern

Attempts to account for the PR phenomenon fall into three main categories, each of which we outline in the subsections below.

2.1. Intransitive Preferences

One possible hypothesis is that the classic PR pattern reflects preferences that allow systematic intransitivity over binary choices. Let us write the \$-bet as \$, write the P-bet as P, and denote strict preference by $\succ$. Suppose we can find some sure amount of money M such that an individual has the preferences $\$ \succ M$, $M \succ P$, and $P \succ \$$: that is, suppose that for this particular $\{\$, P, M\}$ triple, she has intransitive preferences.

If asked to give her certainty equivalent for the \$-bet, CE(\$), she will state some CE(\$) $> M$; if asked to give her certainty equivalent for the P-bet, she will state CE(P) $< M$. Hence, when asked to give the two valuations and also make a direct choice between \$ and P, as in standard PR experiments, she will report CE(\$) > CE(P) and $P \succ \$$, thereby exhibiting the classic PR pattern. In so doing, there is no bias or error in her responses; she is accurately reporting her preferences but happens to have preferences that, in this case, do not conform with transitivity. Although many decision theories take transitivity as axiomatic, not all theories do so; for example, regret theory (Bell 1982; Loomes and Sugden 1982, 1983) allows preferences of the kind that would produce the classic PR pattern for at least some $\{\$, P\}$ pairs.[2]

Of course, if this intransitivity explanation is correct, it should be possible to find evidence of the $\$ \succ M$, $M \succ P$, and $P \succ \$$ cycles that would underpin the classic PR pattern. A paper by Loomes et al. (1991) reported an experiment where choice cycles in this classic direction did indeed outnumber those in the opposite direction, seemingly to a significant extent. However, only a minority of all responses were cyclical, and it was suggested by Sopher and Gigliotti (1993) that the patterns reported by Loomes et al. (1991) could possibly have arisen purely as a result of noise/error. This is an issue to which we shall return in Sections 2.3 and 3.

2.2. Procedural Biases

A rather different kind of explanation of the PR phenomenon was offered by Lichtenstein and Slovic (1971). Other variants have been suggested since, but what they have in common is the idea that people may not have highly articulated underlying preferences that are always accurately and consistently expressed in response to every task, but rather may to some extent construct their responses according to the nature of the task presented to them and may therefore be systematically influenced by certain features of the different procedures used.

So, for example, when asked to give a certainty equivalent—that is, when asked to give a response in terms of a sum of money—it may be that individuals are (subconsciously) prompted to pay extra attention to the money dimension and underweight the probability information. Perhaps, especially when the task is framed as selling, they initially "anchor" on the bet's most desirable payoff and then arrive at their valuation by adjusting down from that payoff to allow for the possibility that some lower payoff may result; but

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

they do not adjust sufficiently and hence tend to generate a higher certainty equivalent for the bet with the higher payoff, the \$-bet. By contrast, when asked to make a straight choice between \$ and P, it may be that more attention is paid to the chances of winning at least something; since the P-bet offers a greater chance of winning something than the \$-bet, this serves to increase the likelihood that the P-bet will be chosen. In short, the weight given to different dimensions may be influenced by the nature of the elicitation procedure, resulting in a systematic disparity between the preference inferred from direct choice and the preference inferred from the two separate valuations.

Tversky et al. (1990) conducted some experiments intended to try to diagnose the causes of preference reversals, both for risky options and for intertemporal decisions. With respect to \$-bets and P-bets (which they relabelled L and H, respectively), they concluded that the phenomenon was primarily due to what they regarded as overvaluing the L(\$) bet, and was partly due to undervaluing the $H(P)$ bet, with perhaps 5%–10% of the effect caused by intransitivity.

However, there are reasons to be cautious about the basis for this diagnosis. One reservation concerns the mechanism used to elicit the values of the bets. Participants were told that at the end of any session involving real payoffs, a pair of lotteries would be selected at random. There was then a 50% chance that a participant would be paid on the basis of playing out whichever option he had picked in the direct choice task, and there was a 50% chance that he would instead play out whichever option he had placed a higher value upon. Whereas Tversky et al. (1990, p. 207) argued that this "ordinal payoff scheme" avoided some of the objections that had been made against the standard Becker-DeGroot-Marschak mechanism (Becker et al. 1964), it had the disadvantage that it was no longer necessary to identify the true indifference value for each bet, since it was only the ordering of the values rather than their precise magnitudes that mattered. In the absence of a mechanism designed to give accurate *magnitudes*, it is arguably unsafe to make statements about "overvaluing" or "undervaluing" on the basis of *ordinal* responses.

Moreover, the extent to which one can detect intransitivity depends on whether the parameters selected by the experimenters happen to fall within a participant's "critical" region (since regret theory, for example, does not entail preferences that are *everywhere* intransitive but only allows that they may be intransitive over some range). Therefore, an experiment that uses a small set of options involving a limited number of preset parameters may well hit upon the critical region for *some* individuals but might miss it for others with different underlying preferences, even though these people might display intransitivity in other (unexplored) triples.

The studies conducted by Bostic et al. (1990) tried to address the latter issue by eliciting choices via two different iterative procedures (while using the Becker-DeGroot-Marschak (1964) mechanism to incentivise valuations elicited in a more conventional way). Bostic et al. (1990) found that their first iterative choice experiment reduced the prevalence of cycles compared with classic PR patterns but did not eliminate the asymmetries between cycles for two $\{\$, P\}$ pairs out of four. Their second experiment, using a more concealed iterative choice procedure,[3] seemed to reduce the significance of asymmetrical cycles even further but had a rather small sample of just 21 respondents.

2.3. Imprecise or Probabilistic Preferences

There is a third kind of explanation that focuses on the possible importance of the *imprecise* or *probabilistic* nature of many people's preferences.

At the less structured end of the spectrum of such accounts, MacCrimmon and Smith (1986) suggested that the classic PR pattern could be explained in terms of the \$-bet allowing a much wider range of valuation responses that do not violate first-order stochastic dominance than the range allowed by the P-bet, so that individuals who were unsure about their precise certainty equivalent could more easily pick higher values for the \$-bet than for the P-bet without being obviously wrong. If both bets have (roughly) the same minimum payoff (small negative amounts in the early experiments, often zero in more recent experiments), there is more scope for giving higher values for the \$-bet than for the P-bet but less scope for giving lower values, which would be sufficient to produce the classic asymmetry.

MacCrimmon and Smith (1986) noted that if the same reasoning were applied to the elicitation of probability equivalents, the opposite asymmetry might be expected.[4] Butler and Loomes (2007) conducted an experiment to explore these possibilities and found some evidence that people's uncertainty about their own preferences varied in ways consistent with MacCrimmon and Smith's conjectures.

At the more structured end of the spectrum, Blavatskyy (2009) proposed a model that embeds a deterministic expected utility theory (EUT) core in a particular stochastic specification and has shown how *some* asymmetry in the classic PR direction might result. The detail of this approach is different from that used by Sopher and Gigliotti (1993) mentioned above, but the general proposition is similar: namely, that if an individual's expressed preferences involve some stochastic component, that component, although itself random, may interact with core preferences in such a way as to produce seemingly systematic departures from standard presumptions. Under certain conditions, this model can even accommodate what Fishburn (1988, pp. 45–46) called "strong" reversals.[5]

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Johnson and Busemeyer (2005) explored the possibility that Busemeyer and Townsend's (1993) decision field theory might provide an explanation. The key idea is that individuals arrive at a valuation response after a cognitive process of iteration between each bet and some sequence of sure amounts. They hypothesised that the starting point of such a process typically involved higher sure amounts when \$-bets were being valued than when P-bets were being processed and that it was this that tended to produce CE(\$) > CE(P) even when $P >$ \$ in a direct comparison.

In Section 5, we will discuss this and the various other possible explanations outlined above. First, however, we set the scene for our experiments.

3. A Broad Probabilistic Choice Framework

It has often been observed that when the same individual is presented with exactly the same tasks framed in exactly the same way on more than one occasion within a fairly short period of time, the individual may answer somewhat differently in at least some of those repetitions. An early manifestation of such behaviour was reported by Mosteller and Nogee (1951), who encountered variability of this kind when they presented respondents with a variety of choices, each repeated multiple times over a period of several weeks. For example, one series of decisions asked respondents either to accept or refuse a gamble that involved a $\frac{1}{3}$ chance of losing 5 cents and a $\frac{2}{3}$ chance of winning X, where X took a number of different values ranging from 5 cents to 16 cents. Over the course of the experiment, each level of X was presented to each respondent on up to 14 independent occasions. Figure 2 of Mosteller and Nogee (1951) depicted a respondent who never accepted the gamble when X was 5 cents or 7 cents and always played it when X was 16 cents, but for intermediate values, his acceptance rate lay between 7% and 93%, increasing monotonically with X. Such variability (although often less neatly monotonic) was typical of the other participants in their study and has been observed in many subsequent studies involving repeated choices.

Over the years, such variability has been formally modelled in a number of ways; see Luce and Suppes (1965) for an early review and Rieskamp et al. (2006) for a more recent one. However, as Stott (2006) and Blavatskyy and Pogrebna (2010) have shown, different assumptions about the way in which the stochastic component is specified can produce quite different estimates of underlying parameters. To avoid becoming embroiled in debates about the sensitivity of our results to particular functional forms, our strategy in this paper is to try to investigate the issues raised above within a framework of probabilistic choice so general that it encompasses a very broad range of more specific stochastic models and relies on a bare minimum of assumptions.

Consider a case where an individual is presented with a number of choices between some lottery B and a series of increasing sure amounts denoted by A_j. If these choices could be presented on a number of different occasions and under circumstances where an individual makes each choice independently of every other one—in the sense of not remembering previous choices and therefore making each new choice afresh—then we might model an individual's underlying preferences as a probability distribution.

Figure 1 shows two possible distributions. The dashed line shows the deterministic preferences of an individual who is indifferent between B and a sure amount A_6: for all $j < 6$, she always chooses B; for all $j > 6$, she never chooses B; and for $j = 6$, all probability mixes of A and B are equally good.

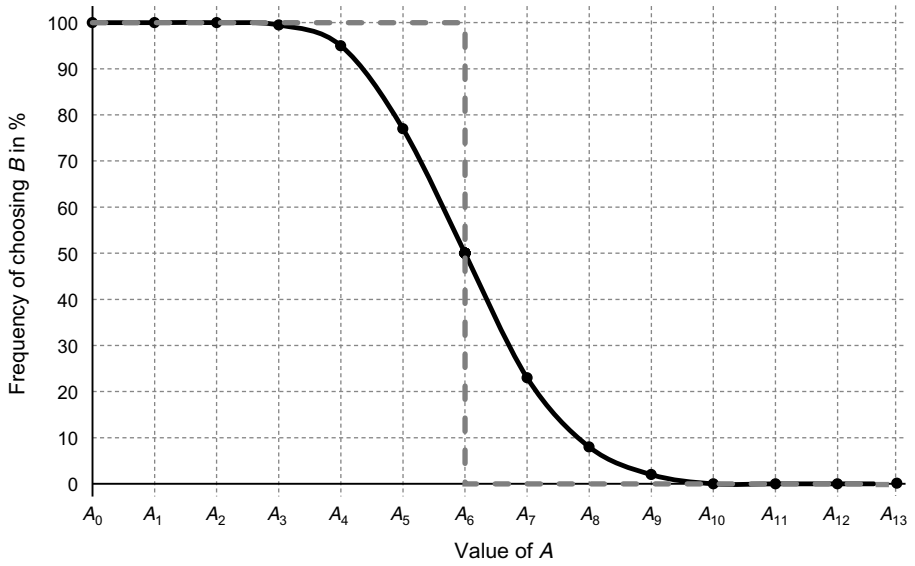
The solid curve shows the underlying distribution for an individual with probabilistic preferences. In this example, when the sure thing is less than or equal to A_3, the individual always chooses B; when the sure thing is equal to or greater than A_{10}, he never chooses B. But for values of A between A_3 and A_{10}, there is some chance that either A or B might be chosen. For example, when presented with a choice between B and A_5, there is a 0.77 chance he will choose B; therefore, if he made that choice 100 times on separate and independent occasions, we would expect on average to observe him choosing the sure amount on 23 of those occasions. Increasing the sure amount to A_8 increases the likelihood that the sure sum will be chosen—to 0.93—but in 100 repetitions of the choice we would still expect B to be chosen on average on seven occasions.

Using $\Pr(A \succ B)$ to denote the probability of choosing A from the pair $\{A, B\}$, we shall refer to cases where $\Pr(A \succ B) = \Pr(B \succ A) = 0.5$ as cases of *stochastic indifference* (SI). This is the stochastic analogue of the notion of certainty equivalence in deterministic theories. In the example shown in Figure 1, the SI point is at A_6.

The curve in Figure 1 is no more than an illustration, and we make no strong claims about its shape. We have drawn it as sigmoid because that seems to fit many intuitions and data sets. We leave open the question of symmetry. Some model specifications may imply symmetry, whereas others may suggest particular kinds of asymmetry; however, such detail is not necessary for our purposes. All we wish to convey in Figure 1 is the distinction between deterministic and probabilistic choice and the broad proposition that models of probabilistic choice allow there to be some range between the point where B is always chosen and the point where B is never chosen,[6] and that between those two points the probability of choosing B falls monotonically as A is unambiguously improved.

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Figure 1. Deterministic and Probabilistic Preferences

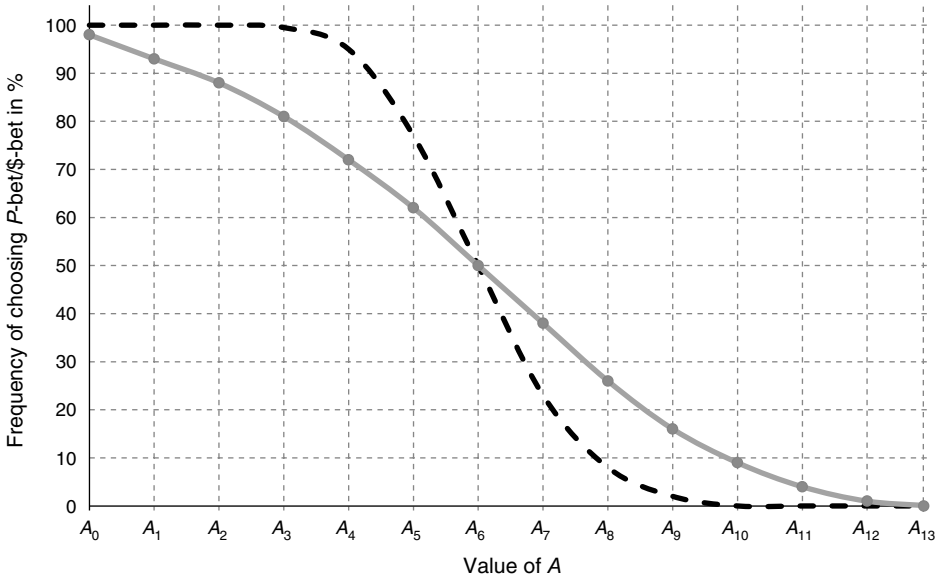


Now let us consider two different lotteries represented on the same diagram. In Figure 2, let the dashed curve represent the distribution for a particular P-bet, and let the solid curve represent the distribution for a particular \$-bet, where $\Pr(\$ \succ P) = \Pr(P \succ \$) = 0.5$ and where both of these bets are stochastically indifferent to the sure amount A_6. Thus, we have constructed a case that is consistent with weak stochastic transitivity (WST), which, for any triple $\{X, Y, Z\}$, requires that if $\Pr(X \succ Y) \geq 0.5$ and $\Pr(Y \succ Z) \geq 0.5$, then $\Pr(X \succ Z) \geq 0.5$.

However, even when WST holds, patterns of choice over specific pairs located away from the SI points may exhibit what looks like systematic disparities. To see this, suppose that we have a sample of 100 individuals whose preferences are each as in Figure 2. Instead of asking them to choose between pairs involving A_6, suppose we set the sure amount at A_8.

Each participant is assumed to make each choice as if on the basis of independent draws from the underlying distributions in Figure 2. Most will give responses that are consistent with transitive orderings over $\{\$, P, A_8\}$, but there will be some sets of choices that are intransitive, either in the direction consistent with the classic PR or in the opposite direction.

Figure 2. Two Different Lotteries Having the Same SI When Compared with A



Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

The cycle corresponding with the classic PR pattern is $\$ \succ A_8$, $A_8 \succ P$, and $P \succ \$$. From Figure 2 we read off $\Pr(\$ \succ A_8) = 0.27$ and $\Pr(A_8 \succ P) = 0.93$. Combining these probabilities with $\Pr(P \succ \$) = 0.5$ and assuming independent draws, the product of these probabilities is 0.12555, so that on average we should expect to observe 12 or 13 members of our sample exhibiting the cycle consistent with the classic PR pattern. The opposite cycle involves $\$ \succ P$, $P \succ A_8$, and $A_8 \succ \$$, for which the probabilities are 0.5, 0.07, and 0.73, respectively, giving a product of 0.02555, so that we should expect two or three people to report this pattern. Even if we round the first number down and round the second number up, we have a 12:3 ratio between the two cycles. If we were to apply the standard binomial test, in effect, testing the null hypothesis that both types of cycles are equally likely, we should reject that hypothesis at the 5% level. Thus, we can see how probabilistic preferences, which respect transitivity at the core, may nevertheless give the *appearance* of systematic asymmetries in responses to a particular set of predetermined options.

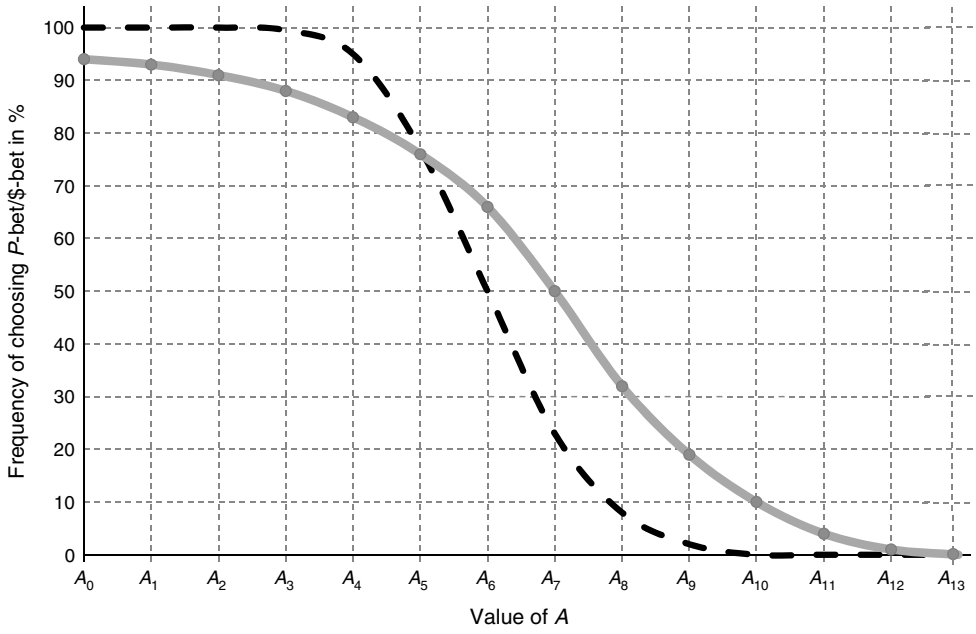
How might we depict probabilistic preferences that do *not* respect WST? Suppose we could find a $\{\$, P\}$ pair such that $\Pr(P \succ \$) > 0.5$ while at the same time the underlying probability distributions for each bet against different sure sums is as in Figure 3.

As in Figure 2, the dashed curve represents the $\{A_j, P\}$ relationship, and the solid curve shows $\{A_j, \$\}$. Here, the SI for the \$-bet is A_7, whereas the SI for the P-bet is A_6. So if we set the sure sum between A_6 and A_7—halfway between, for example, which we denote by A^*—then we should see WST violated because at that level of A, $\Pr(\$ \succ A^*) = 0.58$ while $\Pr(A^* \succ P) = 0.65$, which should imply $\Pr(\$ \succ P) \geq 0.5$, whereas we have a $\{\$, P\}$ pair such that $\Pr(\$ \succ P) < 0.5$.

Notice, however, that such a violation of WST is only observable for A_j between A_6 and A_7. If we were to set the sure sum at anything less than A_6, both $\Pr(\$ \succ A_j)$ and $\Pr(P \succ A_j) > 0.5$, so that no violation of WST can be detected and likewise when the sure sum is greater than A_7 so that both of those probabilities are less than 0.5. Even if underlying preferences are as shown in Figure 3, the scope for demonstrating violations of WST in the direction consistent with the classic PR pattern is limited. It is easy to see that even if all individuals have preferences with the potential to violate WST, they may have different SIs such that any particular preset value of A will only fall into the critical range for a subset of them.

So if we wish to provide a comprehensive examination of the extent to which probabilistic preferences respect WST for \$- and P-bets, we need to elicit *from each individual* some estimate of his or her SI for each lottery against some set of sure amounts and see how the ordering of these SIs compares with the modal choice between them. Denoting the SI for the S-bet as SI(\$) and the SI for the P-bet as SI(P), the possibilities are as follows. If $\text{SI}(P) \geq \text{SI}(\$)$ and $\Pr(P \succ \$) \geq 0.5$, or if $\text{SI}(\$) \geq \text{SI}(P)$ and $\Pr(\$ \succ P) \geq 0.5$, the individual conforms with WST. However, if $\text{SI}(\$) > \text{SI}(P)$ but $\Pr(P \succ \$) > 0.5$, the individual violates WST in the direction consistent with the classic PR pattern, whereas if $\text{SI}(P) > \text{SI}(\$)$ but $\Pr(\$ \succ P) > 0.5$, the violation of WST is in the opposite direction. The experimen-

Figure 3. \$-bet SI > P-bet SI



Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

tal design described in Section 4 seeks to obtain the required estimates of SI(\$), SI(P), and Pr(\$ > P).

4. Experimental Design

4.1. General Principles

In principle, we should like to take various \$-bets and P-bets and, for each bet, identify the range of sure sums that cover the distance between the highest sure sum, which is (almost) never preferred to the bet, and the lowest sure sum, which is (almost) always preferred to the bet. Within that range, we would like to have identified enough pairs, each repeated enough times (ideally, with each choice independent of all earlier presentations of the same pair) to provide a good estimate of the curve in question and a reasonably precise estimate of the SI point.

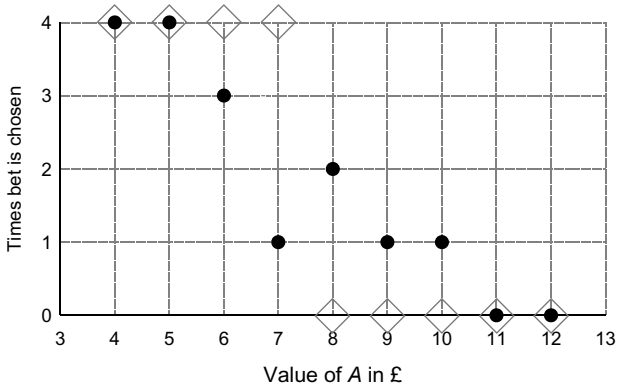
Bearing in mind that any sample is likely to involve some degree of heterogeneity between individuals, in an ideal world one might like to have different sure sums £0.50 or £1 apart covering most of the range between the upper and lower P-bet payoffs, with perhaps 10 repetitions of each $\{A_j, \$\}$ and $\{A_j, P\}$ pair. One might also like to have comparable numbers of repetitions between each $\{\$, P\}$ pair. However, such a design could easily entail four or five hundred choices for a single $\{\$, P\}$ pair, which, when interspersed among a similar number of "distractor" choices, would be a daunting prospect for potential participants and might compromise the quality of the data.

On the basis of our own and colleagues' experience and a pilot study, we judged that we should produce a design using no more than 200–300 questions in total, with a number of these questions involving tasks other than binary choices to try to provide some variety and extra interest for participants.[7] The net result was a design that included up to 36 different pairs, with each pair presented as binary choices (BCs) on four occasions during the experimental session and with each of these presentations being separated from one another by a number of other pairs also presented in random order.

We shall set out in subsequent subsections the particular parameters used, and we shall present evidence to demonstrate that the data showed reasonable sensitivity to variations in those parameters. But first we address a possible concern about whether four repetitions are sufficient for our purposes. Certainly, if our objective were to produce a reasonably tight-fitting curve for each individual for each lottery, four observations per pair would not be sufficient. However, for the purpose of covering the range within which nearly everyone's SI is located and thereby getting an estimate of each SI, we believe that our design is adequate.

To explain how we analysed the data, consider Figure 4, which shows on the horizontal axis the different sure values that A_j might take. The vertical axis

Figure 4. Estimating SIs for Two Bets



measures the numbers of times some bet is chosen in any four repeated pairings with the same A_j.

The black dots represent the responses of an individual with probabilistic preferences. When asked to choose between a \$-bet and nine sure sums, each presented on four separated occasions, the individual always chooses the \$-bet when the alternative is £4 or £5 and never chooses it when A_j offers £11 or £12 for sure. But for sums from £6 to £10, she chooses each alternative on at least one occasion—and is even observed to choose the \$-bet more often (twice) when the sure amount is £8 than when the sure amount is £7. Of course, such seeming inconsistency is likely to occur simply by chance if the individual is behaving as if sampling just a few times from an underlying probability distribution of the kind described in Section 3.

The individual's choices between a P-bet and the same set of sure amounts are shown by the grey diamonds. She always chooses the P-bet when the sure alternative is £7 or less, and she never chooses the bet when the sure sum is £8 or more. Her behaviour here is indistinguishable from what we might expect of someone with deterministic preferences whose CE lies between £7 and £8, and if we have to give a best estimate of that CE, we have no reason to do other than take the midpoint of £7.50. The same applies to our best estimate for the SI of an individual with probabilistic preferences who reports those choices.

Where does this individual's sure-sum SI for the \$-bet stand in relation to her SI for the P-bet? There is more variability in the \$-bet responses, but if we had to judge which of the two bets had the greater SI in terms of sure sums, we should have to conclude on this evidence that there is no significant difference between them: for £4, £5, £11, and £12, the same choices are made throughout; when the sure sum is £6 or £7, the P-bet is chosen four more times; but when the sure sum is £8, £9, or £10, the P-bet is chosen four fewer times. With the small number of discrete choice observations involved, sophisticated econometric estimation may offer little more than we can achieve by simply

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

counting the number of times each bet is chosen out of the total of 36 decisions. In the above example, each bet is chosen 16 times against the same set of sure amounts, mapping to an SI value of £7.50.

Had the individual in the above example chosen, say, the P-bet more often, it would have suggested a higher SI value for that bet. For example, suppose that she had made the same choices as above except that she had chosen the P-bet twice in the $\{£8, P\}$ pair. On that basis, an SI of £8 would seem to be the best estimate. In other words, choosing the bet 18 times in total maps to an SI value of £8. More generally, when the range of sure options is as shown in Figure 4 and when each $\{A_j, \text{bet}\}$ pair is presented on four separated occasions, we estimate SI(bet) = £3.5 + 0.25B, where $0 < B < 36$ is the total number of times a particular bet is chosen.[8]

We shall therefore conduct our analysis of SI points in terms of the numbers of times the bet in question is chosen from any given range of alternatives. There will, of course, be some sampling error for such a measure, but our sample sizes in conjunction with the within-subject nature of the analysis will still allow us to draw a number of conclusions.

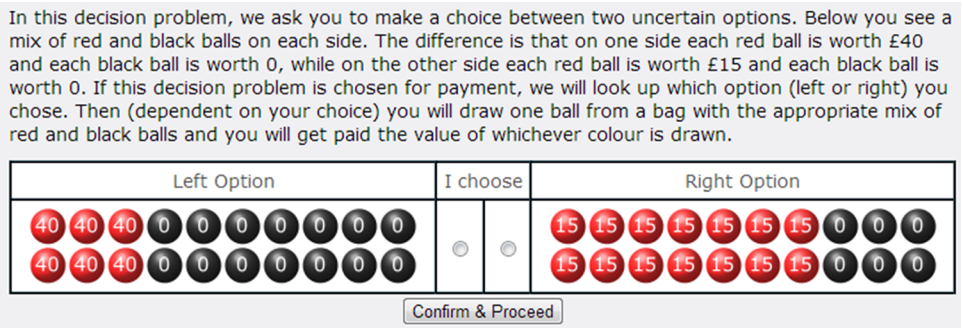
4.2. Experiment 1

Experiment 1 examined the relationship between direct choice between \$ and P and the ordering of the SI values inferred from repeated choices involving the same set of sure amounts. In short, it was designed to look for evidence about conformity with—or else systematic departure from—WST when such bets are involved. The conventional wisdom, as expressed in Rieskamp et al. (2006, p. 648) is that violations of WST are quite rare and tend to occur only in "fairly unusual" circumstances, so that "the principle of weak stochastic transitivity should generally be retained as a bound of rationality." Unfortunately, a number of the experiments they cite could be argued to have involved too few repetitions over too narrow a range to provide a really strong foundation for this conclusion. Our experiment had the drawback that it was constructed around just one $\{\$, P\}$ pair, but its strength was that it involved multiple repetitions and therefore gave a reasonable chance of detecting violations of WST, if there were any to be detected.

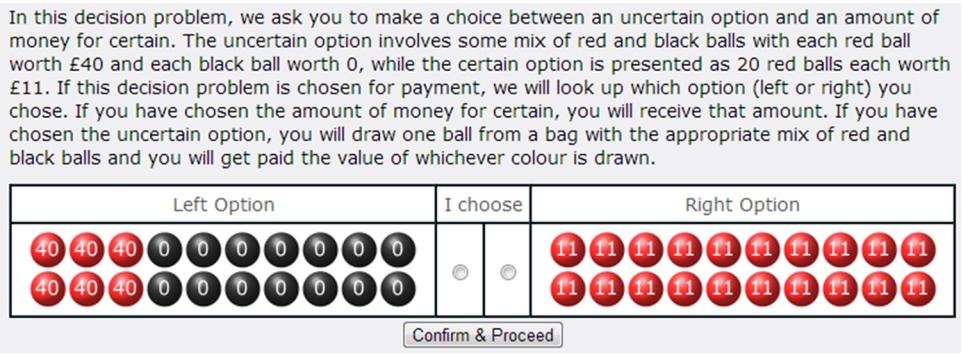
In our pair, what we shall call the *focal* \$-bet offered a 0.3 chance of £40 (otherwise 0), and our *focal* P-bet offered a 0.7 chance of £15 (otherwise 0). Each bet was displayed as shown in the upper panel of Figure 5 when participants were being asked to make a straight choice between the two bets and as in the lower panel of Figure 5 when the alternative was a sure sum of money—in this case, the choice is between our focal \$-bet and £11 for sure. The text accompanying each kind of choice is also reproduced.

Figure 5. (Color online) Examples of Displays Used in Binary Choices

(a) Binary choice between the focal \$-bet and the focal *P*-bet



(b) Binary choice between the focal \$-bet and an amount of money for certain



Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

We opted for this way of displaying alternatives to try to strike a compromise between the "decision by description" and "decision by experience" approaches. A growing literature[9] suggests that when people form an estimate of probabilities on the basis of some sampling experience, they may behave differently from when the probabilities are merely described in decimal or percentage form without the opportunity for participants to get some "feel" for them. The large number of decisions in our design made it impossible to ask people to learn the probabilities for each choice by sampling, but by depicting the distributions of balls that give positive or zero payoffs in a format that allowed probabilities to be easily seen and compared, we hoped to provide a visual proxy for experience by showing exactly what each option would involve in terms of the 20 balls that would be put into a bag when one of the choices came to be played out for real.

The focal \$-bet was presented on four separated occasions in choices with nine different sure integer amounts from £4 to £12, providing a total of 36 BCs. The focal P-bet was offered against the same nine sure sums, giving another 36 choices. We also had nine pairs where the focal \$-bet was fixed while the alternative bet's probability of £15 varied from 0.5 to 0.9 with increments of 0.05, each presented on four separated occasions; and another nine pairs where the focal P-bet was held constant while the alternative bet's probability of £40 varied from 0.15 to 0.55 with increments of 0.05. Thus, we have a total of $4 \times 9 \times 4 = 144$ BCs. For each individual, then, we can estimate a certainty equivalent SI for the focal \$-bet and a certainty equivalent SI for the focal P-bet, and we can compare the ordering of these SIs with the frequency of choice from a total of eight separated presentations of the focal $\{\$, P\}$ pair.

This experiment involved 101 participants through the online recruitment system of the Decision Research at Warwick (DR@W) Group in the University of Warwick. Each participant received an invitation with detailed instructions together with a link to the online experiment. Participants were invited to complete the online experiment in their own time by a specified deadline. Each invited participant was assigned a unique ID number, which was automatically copied as a password to the experimental interface. This ensured that (a) only invited participants could take part in the experiment, and (b) none of participants could take part more than once. The experiment was computerised using the Experimental Toolbox (Expert)[10] online platform.[11]

Besides the 144 BC questions described above, there were a further 36 questions presented in the form of four nine-row choice lists; these took the total of incentivised lottery-based tasks to 180, with these being mixed up and spread over sections 1, 3, and 5 of the experiment. Sections 2 and 4 consisted of quite different tasks involving 2×6 hypothetical questions of the type used in the domain specific risk attitude (DoSpeRT) procedure (Blais and Weber 2006). These two sections were included as "distractor" tasks for our participants, providing greater separation between repetitions of the incentivised questions. They play no part in our analysis.

The incentive mechanism was as follows. On the date of the specified deadline, 100 participants were selected at random from all participants who completed the experiment on time.[12] These participants were invited to the DR@W experimental laboratory for individual scheduled appointments. One of the 180 incentive-linked questions was picked at random and independently for each participant and was played out for real money. There was no show-up fee, and the instructions made it clear that the participant's entire payment depended on how her decision played out in the one randomly selected question.

If the participant had chosen some sure amount of money, she would simply receive that sum. If she had chosen a lottery, she would see an opaque bag being filled with the numbers of red and black balls (actually, coloured marbles) specified in the question. She then picked a marble at random and was paid (or not) accordingly.

This incentive mechanism was described to all participants in the instructions. Participants also received a practice question and had an opportunity to email the experimental team in case they were not clear about the instructions. On average, it took each participant 30–40 minutes to complete the online experiment. Each individual appointment in the DR@W laboratory lasted between three and five minutes. The average payoff in the experiment was approximately £12.

We begin by presenting various aggregate patterns of response, so that readers can form some view about the extent to which the data showed broad regularities of the kind most models might entail. Table 1 shows how the likelihoods of choosing the fixed lotteries changed as the alternatives were progressively improved. In all cases, the proportions were sensitive in the expected directions to variations in parameters.

At this aggregate level, the central tendencies look broadly consistent with transitivity. Overall, the P-bet was chosen in just over 83% of the direct choices between the focal \$-bet and the focal P-bet. At the same time, on the basis of the choices between each bet and the different sure amounts, the sample median valuation of P was a little over £9 and the sample median valuation of \$ was £7. At this level of analysis, the majority choice was compatible with the ordering of median values.

However, the usual PR asymmetry is an individual-level phenomenon. Therefore, Table 2 assigns individuals to cells according to the relationship between the

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Table 1. Aggregate Binary Choice Proportions in Experiment 1

	Sure amounts								
	£4	£5	£6	£7	£8	£9	£10	£11	£12
$-bet[a] chosen %	78.7	65.4	60.2	50.0	41.6	35.9	15.8	15.6	8.7
P-bet[a] chosen %	92.1	87.4	81.2	75.2	64.9	53.2	28.0	18.8	12.9
	Chances of £15 offered by variants of the *P*-bet								
	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
$-bet[a] chosen %	59.9	34.9	29.5	18.3	17.6	12.6	7.2	6.4	4.2
	Chances of £40 offered by variants of the $-bet								
	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
P-bet[a] chosen %	98.5	97.0	94.3	84.4	74.8	58.2	43.3	26.0	17.8

[a]Focal $-bet = (£40, 0.3; 0, 0.7); focal *P*-bet = (£15, 0.7; 0, 0.3).

frequency with which they chose $ or *P* in the direct choice between them (shown in the rows) and the ordering over the two bets inferred from the difference between their SI certainty equivalents, with the SI differences being grouped into seven column categories.

The observations that represent a strict reversal of one kind or the other are shown in bold. In total, 10 of the 101 individuals fell into these cells, with another 10 either choosing each bet equally frequently or having the same SI certainty equivalents (or both, in one case). The 10 strict reversals were divided 9:1 in the "usual" direction—an asymmetry that appears unlikely to have occurred by chance—but by comparison with most PR experiments involving direct elicitation of selling/buying prices, the overall proportion of reversals was low. Moreover, the magnitudes of the differences were small by comparison with many selling price experiments—in the present experiment, the greatest SI difference in the "wrong" direction was just £1.50. So it might be argued that there was *some* tendency toward intransitivity in the classic direction that cannot be explained simply in terms of noise/error; but the effects were modest and the overall picture is compatible with the great majority of participants' responses respecting weak stochastic transitivity.

However, Experiment 1 was built around just one particular $\{\$, P\}$ pair, and it could be objected that the results might not generalise to other pairs. Moreover, Experiment 1 focused exclusively on choice-based tasks and did not at the same time elicit valuations in a more direct way, using the kind of incentive mechanism more commonly used in conjunction with selling prices in traditional preference reversal experiments. Therefore, in Experiment 2 we used slightly more extreme bets that might be regarded as (even) more typical of many PR experiments, and we added a conventional valuation task to the other formats used in Experiment 1 so that we could make direct within-person comparisons.

4.3. Experiment 2

Many features of this experiment were the same as in Experiment 1 in terms of the numbers of binary choices and the incentive system for those choices. The two key differences between this experiment and the pre-

Table 2. Choice and SI Valuation Differences Inferred from BCs in Experiment 1

	SI(*P*) − SI($)							
	≥ £1.75	£1.50 to £0.75	£0.50 to £0.25	0	−£0.50 to −£0.25	−£0.75 to −£1.50	≤ −£1.75	Total
Frequency of choice								
8 *P*; 0 $	33	15	13	4	**2**	**3**	—	70
7 *P*; 1 $	2	1	2	—	**2**	**1**	—	8
6 *P*; 2 $	—	2	1	—	—	—	—	3
5 *P*; 3 $	1	2	—	—	—	**1**	—	4
4 *P*; 4 $	1	—	—	1	—	—	—	2
3 *P*; 5 $	—	—	—	1	—	—	1	2
2 *P*; 6 $	—	—	—	—	—	2	—	2
1 *P*; 7 $	—	—	**1**	1	—	—	—	2
0 *P*; 8 $	—	—	—	2	—	5	1	8
Total	37	20	17	9	4	12	2	101

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

vious one were as follows. First, the new focal $-bet offered a 0.25 chance of £50 (otherwise 0), and the new focal *P*-bet offered a 0.8 chance of £12—that is, the two bets were a little farther apart in terms of probabilities and payoffs and expected values than the pair in the first experiment. Second, we dropped the DoSpeRT questions and replaced them in sections 2 and 4 with 20 direct valuation (DV) tasks linked to the kind of incentive mechanism often used, owing to Becker et al. (1964). Those 20 DV tasks involved five different lotteries, each valued on two separated occasions in section 2 and again twice each in section 4, thus giving four separated valuations for each lottery. Figure 6 shows an example of a typical DV task—in this case, for the focal *P*-bet in the current experiment.

Two of those five lotteries were the $-bet and *P*-bet, which are focal to this experiment. Two others were the $-bet and *P*-bet used in Experiment 1. The fifth lottery was one used in the "independence" experiment run in parallel with Experiment 1, as referred to in Endnote 12.

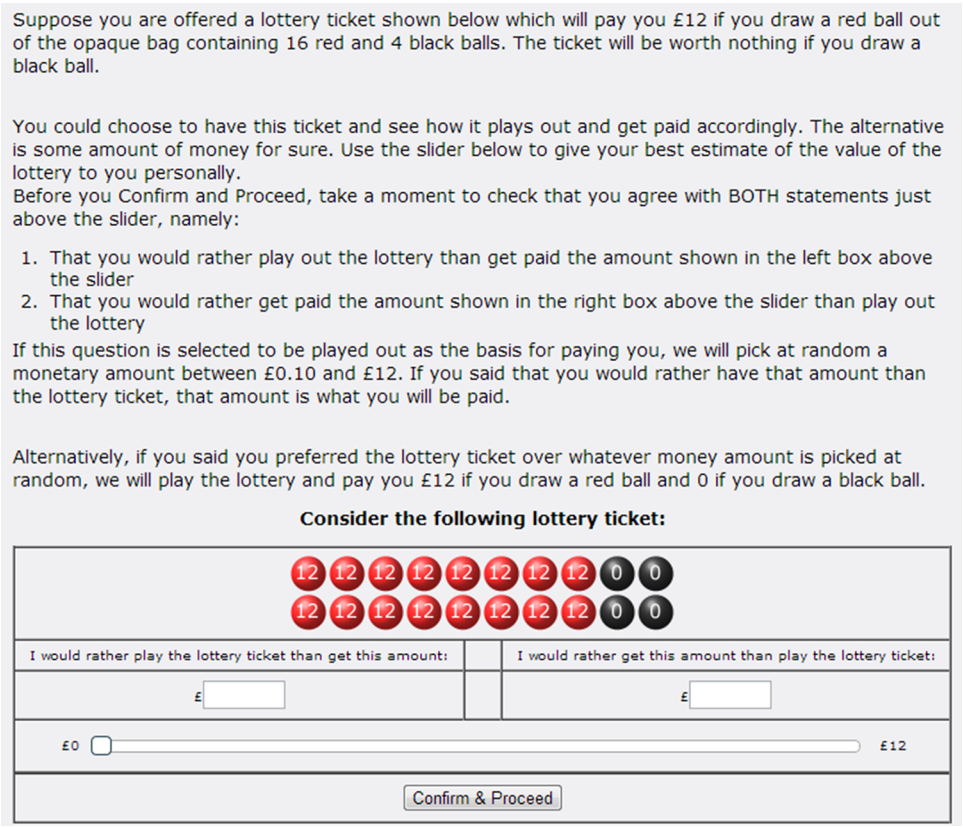
By embedding the DV tasks among many BCs involving the same or similar bets and sure amounts, our intention was to give participants every opportunity to be consistent across the two types of tasks. We also tried to formulate the valuation task to be as much like a choice task as we could, avoiding any reference to "price" or to selling or buying. That is, we were consciously attempting to reduce any framing effects so as to examine the question of whether a valuation task per se was treated differently than a series of binary choices.

To respond, participants used a mouse to click on the button—which was always initially located at zero to avoid any differential "starting point" effects between the different lotteries—and to move it along the slider. As the button moved, two values that were multiples of £0.10 and were always £0.10 apart appeared in the boxes just above the slider, rising or falling as the button was dragged right or left. Once a participant had settled on a response, he clicked on "Confirm & Proceed" and then either valued a fresh lottery or else, at the end of each series of 10 valuations, moved on to another series of BCs.

So, for example, an individual might move the slider, steadily increasing the amounts shown in the boxes until, say, the left-hand box displayed £7.50 and the right-hand box displayed £7.60. The respondent would thereby be stating that he would rather play the lottery than get £7.50 and would rather get £7.60 than play the lottery.

A total of 184 participants from the University of Warwick completed all of the BC and DV questions.

Figure 6. (Color online) Example of Display Used in Experiment 2 Direct Valuation



Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Table 3. Aggregate Binary Choice Proportions in Experiment 2

	Sure amounts							
	£4	£5	£6	£7	£8	£9	£10	£11
\$-bet[a] chosen %	70.7	58.0	47.6	38.7	32.7	26.9	15.5	14.8
P-bet[a] chosen %	92.3	88.5	84.4	69.4	53.8	45.4	20.9	13.9
	Chances of £12 offered by variants of the *P*-bet							
	0.65	0.7	0.75	0.8	0.85	0.9	0.95	
\$-bet[a] chosen %	39.5	31.0	25.3	20.1	17.5	14.0	12.2	
	Chances of £50 offered by variants of the \$-bet							
	0.1	0.15	0.2	0.25	0.3	0.35	0.4	
P-bet[a] chosen %	97.4	97.0	89.4	80.0	70.2	58.3	43.3	

[a]Focal \$-bet = (£50, 0.25; 0, 0.75); focal *P*-bet = (£12, 0.8; 0, 0.2).

On this basis, for each participant, we could identify from the BC responses

- the individual's SI certainty equivalent for the \$-bet,[13]
- the individual's SI certainty equivalent for the *P*-bet, and
- the distribution of eight straight choices between \$ and *P*.

From the 20 DV questions, we had[14]

- four estimates of the certainty equivalent of the \$-bet in this experiment,
- four estimates of the certainty equivalent of the *P*-bet in this experiment,
- four estimates of the certainty equivalent of the \$-bet used in Experiment 1,
- four estimates of the certainty equivalent of the *P*-bet used in Experiment 1, and
- four estimates of the certainty equivalent of a lottery offering (£40, 0.8; 0, 0.2) used in the independence experiment for which we have BC-based SIs.

Although the last three sets of certainty equivalents do not give us the within-person information that we can get from the first two, the participants in all experiments were drawn from the same sampling frame—the DR@W list mentioned earlier—and we hoped that the additional between-sample information might prove useful.

Table 3 presents some aggregate results for the BC responses in the current experiment, comparable with the data presented in Table 1 for Experiment 1.

As before, the aggregate data look responsive in the expected direction across the different parameter variations. The sample median value for the *P*-bet was somewhere between £8 and £9, and the sample median value for the \$-bet was below £6. In direct choices between the focal \$-bet and the focal *P*-bet, *P* was chosen 80% of the time. So, as before, the aggregate data look broadly consistent with transitivity.

However, as before, the data that are most relevant for our purposes are the individual-level comparisons. Table 4 displays the BC-based results for Experiment 2 in a similar format to that used in Table 2 for the previous experiment.

Again, the numbers shown in bold font highlight cases where there were strict reversals between the ordering inferred from the SIs and the majority of choices in the $\{\$, P\}$ pairs. This time there were 13 individuals—just over 7% of the sample—exhibiting such reversals, dividing 12:1 in the direction consistent with the usual PR asymmetry. So although this asymmetry was unlikely to have occurred by chance, the overall magnitude of the effect was again rather modest and far short of the scale necessary to underpin the rates of preference reversal reported in most studies.

We now consider the DV responses generated by the same 184 individuals. Most participants' four DVs for each bet showed some degree of variability: if we compute the standard deviation of each individual's four DVs for each bet, there were just 16 (8.7%) who had zero standard deviations for both bets.[15] The sample average (median) of individual standard deviations was 0.88 (0.58) for the *P*-bet and 4.32 (2.89) for the \$-bet. Since the median DVs show less vulnerability to outliers and are comparable with SIs, Table 5 replaces the SI differences in Table 4 by differences between the median values derived from the DV tasks. However, since individuals' DV differences covered a much wider range than their SI differences, the column specifications are adjusted accordingly.

Clearly, the DV tasks elicited responses that were substantially different from the patterns of values inferred from BCs. Table 4 shows that 46 individuals (25% of the sample) had SI(\$) > SI(*P*), whereas Table 5 shows that for 156 (84.8%) individuals, median DV(\$) was higher than median DV(*P*). As a consequence,

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

Table 4. Choice and SI Valuation Differences Inferred from BCs in Experiment 2

	SI(P) − SI($)							
	> £2.00	£1.25 to £2.00	£0.25 to £1.00	0	−£0.25 to −£1.00	−£1.25 to −£2.00	< −£2.00	Total
8 P; 0 $	68	20	19	11	**7**	**2**	—	127
7 P; 1 $	6	—	—	—	**3**	—	—	9
6 P; 2 $	1	—	4	—	—	—	—	5
5 P; 3 $	—	1	2	—	—	—	—	3
4 P; 4 $	—	—	1	2	2	—	—	5
3 P; 5 $	—	—	—	—	2	3	—	5
2 P; 6 $	—	—	—	—	3	2	—	5
1 P; 7 $	—	—	**1**	—	1	2	4	8
0 P; 8 $	—	—	—	2	4	9	2	17
Total	75	21	27	15	22	18	6	184

the DV tasks produced 117 classic preference reversals, compared with just one individual reversing in the opposite direction.

Going from a ratio of 12:1 BC-based reversals to a ratio of 117:1 DV-based reversals is a remarkable disparity. It is striking that 102 (87.2%) of those 117 chose P on all eight occasions when asked to make a direct $\{\$, P\}$ choice. It is even more striking that no fewer than 85 of these—nearly half of the sample—exhibited strong reversals of the kind referred to in Endnote 5: that is, even though their median value for the $-bet was strictly higher than the £12 payoff offered by the P-bet, they chose the P-bet over the $-bet on every one of the eight occasions when they were asked to make a straight choice. Although Blavatskyy's (2009) model can, in principle, accommodate *some* strong reversals, the parameters we used lie outside the set to which that possibility applies under the conditions he assumed—see Figure 2 in Blavatskyy (2009, p. 246). Such a large and pervasive disparity between BC and DV invites further analysis and discussion.

5. Competing Explanations: Further Exploration of the Data

On the basis of our data, we can confidently reject the proposition that the classic PR phenomenon results *primarily* from intransitive choices. Rather, our data appear to be more in line with the Tversky et al. (1990) suggestion that the phenomenon is mainly due to a mixture of overvaluing the $-bet and/or undervaluing the P-bet.

However, as noted in Section 2, Tversky et al. (1990) only asked a limited range of one-off choice questions, and they elicited valuations via an ordinal scheme that could not be guaranteed to incentivise participants to give their best estimates of magnitudes of values. By contrast, we used a spectrum of repeated binary choices, and in our DV elicitation, we used an incentive mechanism of the more traditional type intended to encourage accurate revelation of actual values. Our data therefore offer opportunities to look more closely at the generality of the Tversky et al. conclusions.

Before doing so, we make a remark about the framing of the conclusions of Tversky et al. (1990). To say that a bet is "overvalued" or "undervalued" might be interpreted as suggesting that there is a "gold standard" against which such values can be judged, with the gold standard in this case presumably being the preferences revealed by binary choice. However, once we begin to think in terms of responses as being probabilistic rather than deterministic, it becomes harder to argue that one type of response reveals "true"

Table 5. Choice and Differences Between Median Direct Values in Experiment 2

	DV(P) − DV($)								
	> £2.00	£0.55 to £2.00	£0.05 to £0.50	0	−£0.05 to −£0.50	−£0.55 to −£2.00	−£2.05 to −£12.00	< −£12.00	Total
8 P; 0 $	12	9	1	3	**6**	**13**	**45**	**38**	127
7 P; 1 $	—	1	—	—	—	—	**4**	**4**	9
6 P; 2 $	—	—	1	—	—	**3**	**3**	—	5
5 P; 3 $	—	—	—	—	—	—	**2**	**1**	3
4 P; 4 $	—	—	—	—	—	1	3	1	5
3 P; 5 $	—	—	—	—	—	—	3	2	5
2 P; 6 $	**1**	—	—	—	—	—	2	2	5
1 P; 7 $	—	—	—	—	—	—	6	2	8
0 P; 8 $	—	—	—	—	—	2	10	5	17
Total	13	10	2	3	6	17	78	55	184

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

preferences, whereas any others that produce different expressions of preference must be to some extent biased. Therefore, we are agnostic on the question of what constitutes "true" preference. However, and subject to that reservation, since most formal decision theories tend to be built on binary relations, we will take the BC responses as our reference point and use the terms "overvalued" and "undervalued" as shorthand ways of describing DVs relative to the BC baseline.

We begin by considering the data from Experiment 2, which allowed us to make 184 comparisons between each individual's median DV and his or her SI point for each bet. Starting with the $-bet, there were just 10 participants for whom median DV($) < SI($) and 1 for whom the two were equal. Thus, 94% of our sample gave a median DV($) strictly higher than the SI($) inferred from their repeated BCs. The sample mean of DV($) based on individual medians was £18.32, compared with a sample mean SI($) of £6.55, giving an average within-person difference of £11.77 between DV($) and SI($). The hypothesis of no difference between the two is strongly rejected ($p < 0.001$). In this respect, our data reinforce the findings of Tversky et al. that the DV task overvalues the $-bet.

However, we do not concur with the conclusion of Tversky et al. (1990) about direct value elicitation undervaluing the P-bet. On the contrary, the DV responses in Experiment 2 also overvalued the P-bet, although to a much lesser extent. There were 57 participants for whom median DV(P) < SI(P), 4 for whom median DV(P) = SI(P) and 123 for whom median DV(P) > SI(P). So two-thirds of the sample gave median DVs than were strictly higher than their SIs. The sample mean DV(P) based on individual medians was £8.87, whereas the sample mean SI(P) was £8.18, giving an average within-person difference of £0.69 and quite firmly rejecting the hypothesis of no difference between the two ($p < 0.01$). Still, the degree of overvaluation was much reduced: the average DV(P) was just 8.4% higher than the average SI(P), compared with the average DV($) being 2.8 times the size of the average SI($). Clearly, the fact that so many people overvalued the $-bet to a much greater extent than they overvalued the P-bet greatly outweighed many of the cases where SI(P) > SI($) and produced the very pronounced classic PR pattern reported in Table 5 and the large number of strong reversals.

Although we cannot conduct the same within-person analysis for the bets used in Experiment 1, we can undertake some between-sample comparisons to provide further relevant evidence.

When asked to make repeated BCs between different sure amounts and $ = (£40, 0.3), the 101 participants in Experiment 1 generated a sample mean SI($) of £7.22. This compares with a sample mean of £15.74 based on the individual median DV($) responses of the 184 participants in Experiment 2—again, more than double the SI-based figure. Not surprisingly, the null hypothesis is strongly rejected ($p < 0.001$).

For $P = (£15, 0.7)$, the differences are smaller, with the sample mean of DV medians being £9.76 compared with an average SI(P) of £8.64 from Experiment 1, so that the DV-based figure is about 12% higher. This difference registers as clearly significant ($p < 0.01$), once again suggesting that direct valuation also overvalues the P-bet relative to binary choice, albeit to a lesser extent than it overvalues the $-bet. Therefore, if we were to infer preferences from median DV responses, we should suppose $ to be preferred to P by 82% of the 184 participants in Experiment 2, whereas on the basis of the SIs in Experiment 1, $ was preferred to P by fewer than 18% of participants. Such a large difference reinforces the suggestion of a substantial disparity between DV- and BC-based preference elicitation methods.

Finally, from another 81 individuals in a separate part of Experiment 1 focusing on the independence axiom, we have SI values for $R = (£40, 0.8; 0, 0.2)$, the fifth bet for which we elicited four DVs from each participant in Experiment 2. The 81 sets of BC responses give a sample mean SI(R) of £25.12, compared with a sample mean DV(R) of £29.02 based on the 184 median DVs. Here, for a bet that can also be regarded as a P-bet, the sample mean of the median DVs is 16% higher, an overvaluation that is again significant at the 0.01 level.

To summarise so far, it seems clear that whatever process generates DV responses is a significantly different process from the one that produces BC responses. The question then is, what can our data tell us about the DV process?

We have seen (in Endnote 15) that with the exception of a small number of individuals who give extreme responses and a similarly small number who apply an expected value rule, the great majority display some variability in their set of four responses per bet. This could be regarded as compatible with DVs being produced by some internal "production" process, which may be different from the process that produces BCs but may exhibit its own within-task regularity.

As mentioned in Section 2.3, one way of modelling such a process has been suggested by Johnson and Busemeyer (2005), applying decision field theory to direct valuation tasks. They propose that individuals can be modelled as arriving at their value responses by means of a squential value matching (SVM) process. In essence, they propose that it is as if an individual starts by comparing a particular lottery with some initial sure amount, then adjusts that sure amount up or down depending on whether the lottery seems clearly preferable or whether the sure amount seems clearly preferable, and repeats this internal iteration until he

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

or she comes to a sure sum such that it is hard to identify a preference for one option or the other, at which point the process is terminated and that sure sum is stated as the value.

In an example they give, Johnson and Busemeyer (2005) take the initial value to be halfway between the two payoffs of a bet. For P-bets this is a relatively low figure (in our Experiment 2, halfway between £12 and 0, i.e., £6), whereas for the $-bet it is much higher (in our Experiment 2, it would be £25). The stochastic nature of the process means that iterations starting low and working up—as for P-bets—are more likely to be terminated at values in the lower part of the interval where there is imprecision, whereas iterations starting high and coming down—as for $-bets—are more likely to end in the higher part. This would produce the kind of result that Tversky et al. (1990) reported, although it would not produce the result we found, where the P-bet was also overvalued somewhat. However, it might seem that we could achieve our result if we modified Johnson and Busemeyer's supposition—which was more of a conjecture than a strong assumption—about the amount used to start the iteration: if people actually start the iteration for both bets from the high payoff and work down, the process could end with the P-bet (coming down from £12) being a little overvalued, whereas the $-bet (coming down from £50) could be much more overvalued. This is more in the spirit of the "anchoring and insufficient adjustment" explanation suggested by Lichtenstein and Slovic (1971, pp. 54–55).

One implication of Johnson and Busemeyer's (2005) model is that the valuation process has much in common with the binary choice process, in the sense that after the initial amount has been picked, the rest of the valuation procedure involves a sequence of binary choices to determine the next step of the iteration. If that were the case, one might suppose that the downward movement from any particular initial value would be least for those with the highest SIs for a particular gamble, since they will tend to have the highest top ends of their imprecision interval. Thus, we should expect that those with the highest SIs should also be the people giving the highest DVs for that same gamble. Therefore, even if the *magnitudes* are rather different between DV and SI, we should still observe a high degree of *rank* correlation between the DVs and SIs for each gamble.

We can examine that hypothesis with respect to both the $-bet and the P-bet in Experiment 2. For the P-bet, the Spearman rank correlation coefficient $\rho = 0.308$ ($p < 0.01$), whereas for the $-bet, $\rho = 0.165$ ($p < 0.05$). Thus the relationship is significant and in the right direction; but each ρ is rather modest, especially in the case of the $-bet.

To provide some comparability, consider an alternative possibility: that the mental process used to generate DV responses is rather different from the mental process used to generate BCs, but each mental process is applied to similar bets in a way that is internally consistent. If *this* were the case, we should expect to see rather higher degrees of correlation between the DVs for the two $-bets and also between the DVs for the three P-bets in Experiment 2 than between DVs and SIs for any particular bet. And this is what we find: for the two $-bets, $\rho = 0.915$ ($p < 0.001$); for the two P-bets, which are the main focus of the two experiments, $\rho = 0.846$ ($p < 0.001$). The third P-bet has an expected value more than three times as large as the other two, but still, the rank correlation is reasonably high: $\rho = 0.725$ ($p < 0.001$) for the comparison with the P-bet from Experiment 1 and $\rho = 0.771$ ($p < 0.001$) for the comparison with the P-bet from Experiment 2.

So there appears to be a high level of regularity within the DV task when applied to similar types of bets. However, as the bets become less similar—i.e., when we compare a $-bet with a P-bet—the correlations become considerably weaker. This would seem to speak against a very simple anchoring and insufficient adjustment model, whereby individuals start from the high payoff and then discount by some factor that is roughly proportional to—but understates—the probability of zero. Were such discount factors to be personal characteristics that vary between individuals but are applied more or less consistently to different bets, we should expect a good degree of rank correlation across $\{\$, P\}$ pairs. There *is* a significant positive rank correlation for these pairs, but it is a good deal lower than between $-bets or between P-bets: for the median DVs for the $\{\$, P\}$ pair from Experiment 1, $\rho = 0.405$ ($p < 0.001$); for the $\{\$, P\}$ pair from Experiment 2, $\rho = 0.282$ ($p < 0.001$).

On the other hand, if we compute for each individual the difference between median DV($) and median DV($P$) in Experiment 2 and compare it with the corresponding within-person difference for the $\{\$, P\}$ pair in Experiment 1, we again find a very high correlation, with $\rho = 0.889$ ($p < 0.001$): that is, those who state the biggest differences in direct valuations between $ and P for one pair are also likely to be reporting the biggest differences for the other pair. At the same time, the ranking of individuals according to their expressed difference between DV($) and DV($P$) is completely uncorrelated with their ranking on the basis of the difference between SI($) and SI($P$), with $\rho = 0.005$ ($p = 0.945$), lending even more force to the conclusion that valuations are not simply some modified form of iterative choice procedure but invoke rather different cognitive processes—although we cannot yet offer an alternative model of those processes that would generate the various patterns reported above.

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

6. Concluding Remarks: Some Challenges for Theory and Practice

Most theories of decision making under risk and uncertainty are presented as deterministic models with general applicability across all types of decision tasks including choice and valuation.[16] To the extent that actual behaviour exhibits some degree of variability, this is often represented in terms of some error term or stochastic specification external to the deterministic core. Such models allow the possibility of seemingly systematic departures from the implications of the deterministic core, but only to a limited degree, as discussed in Section 3.

By asking individuals to make repeated binary choices covering a range of values (and, in Experiment 2, repeated direct valuations), we sought to engage with the probabilistic nature of observed decisions. We thereby obtained estimates of SIs and median DVs that allowed us to examine the extent of within-task and between-task conformity with standard core assumptions.

In two separate experiments using two different, but typical, $\{\$, P\}$ pairs, we found substantial—although not total—conformity with weak stochastic transitivity within the domain of binary choice. To the degree that we found violations of WST, they were systematically in the direction consistent with "classic" preference reversals; but these only accounted for 7%–10% of all participants, and the magnitudes of the disparities were quite small. The usual extent of the preference reversal phenomenon cannot be explained in terms of the choice-valuation analogue of such intransitivity.

Nor can the full extent of the phenomenon be explained by adding external noise to a transitive core that can be applied equally well to choice and valuation. If it were simply a matter of such noise, the effect should be greatly attenuated, if not completely eliminated, by working with the medians extracted from the distributions of response elicited by repeated presentation of each task. However, despite trying to control any obvious framing effects by formulating the DV tasks in a choicelike manner and interspersing the DV tasks among the BC tasks to allow participants every opportunity to process them on a comparable basis, the response patterns were very different and produced a strong pattern of the classic kind.

It is this result that poses the greatest challenge for decision modelling. Whereas it might be true that a few respondents may not fully understand the DV task and/or the implications of the Becker-DeGroot-Marschak incentive mechanism, it is clear that this is not the main source of the disparity. The DV responses exhibited *some* within-person variability, but for the most part, this was modest, and the correlation between DV responses to similar bets was high. Therefore, participants were, for the most part, responding to the DV tasks in ways that showed reasonable *internal* consistency and coherence but that were markedly and systematically different from the ways in which those same people responded to the BC tasks.

One possible explanation is that the variability observed in this and many other experiments is a reflection of cognitive processes that are inherently subject to stochastic effects. In the context of perceptual decisions, such processes have been modelled as sequential sampling/accumulation of evidence (see Otter et al. 2008 for a review). Busemeyer and Townsend's (1993) decision field theory is an early example of such an approach being applied to preference tasks where decisions are seen as the end result of a deliberative process. A possible corollary of this kind of model is that different types of tasks may invoke somewhat different sampling/accumulation processes in ways that generate patterns of response that are reasonably consistent *within* a given task type but do not allow simple inferences to be made from one task type to another. If this is the case, no deterministic theory that claims to apply generally across task type and irrespective of other contextual factors is likely to succeed descriptively. This poses the challenge of developing models that help us to understand procedural effects as part of the fabric of decision making, perhaps enabling us to make appropriate allowances for them.

It also poses a challenge for practical and policy applications. Cost–benefit analyses in areas such as health, safety, or environmental protection, for example, may elicit values for nonmarketed goods through surveys and use such values to guide resource allocation policy in those areas. The most straightforward way of obtaining such values may be to ask directly for individuals' monetary valuations (their "willingness to pay") for the goods in question. However, if the preference orderings inferred from individuals' direct valuation responses do not reflect the choices they would make, we may not be able to rely on such values.

If we want values that more reliably reflect the choices people make, the method used in this study, although suitable for our experimental investigation of probabilistic choice, is not likely to be practicable for general surveys. For such purposes, we would need techniques that can be administered more quickly. However, the speed and convenience of such methods may come at some cost in terms of accuracy or susceptibility to bias. To illustrate the point, we briefly consider three candidate techniques.

Parameter estimation by sequential testing (PEST) is an iterative choice algorithm that was used by Bostic et al. (1990) in their second experiment to see whether preference reversals would diminish or disappear when values were derived from choices. In their first experiment they used a simpler and more transparent iterative procedure to arrive at CEs for each bet.

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

However, for two of the four bet pairs they studied, the usual PR asymmetry persisted to a significant extent. Suspecting that this might be attributable in part to respondents realising that the choice iterations were aimed at eliciting values and thus being prompted to think more in direct valuation terms, Bostic et al. (1990) used the more sophisticated PEST procedure, alternating each step with an unrelated filler question to disguise the algorithm. This less transparent procedure further reduced the PR asymmetry. More recently, the PEST method and another opaque iterative process were used by Sanchez Martinez et al. (2015) in the context of health state evaluation, where "matching" methods led to estimates that diverged significantly from binary choices but where at least some of the concealed iterative methods produced equivalences much more in line with choices. Although further work is necessary to establish more broadly the relative performance of concealed iterative choice procedures compared with the kind of random binary choice method used in our study, it may turn out that some form of disguised iteration could be a feasible option—although it, too, will often entail a nontrivial number of choices, once allowance has been made for the number of steps within each iteration as well as the number of filler tasks required.

Another method that might be used to elicit equivalences involves multiple price/choice lists (MPLs). The general principle here is to construct a table where each row is a binary choice and where the balance between the options on either side of the table changes progressively as we move up or down the list. For example, Cohen et al. (1987) asked respondents to choose between a lottery that was kept fixed on one side of the table, with the options on the other side involving 21 sure amounts of money that increased from one row to the next. The point at which a respondent switched from one side of the table to the other was taken to indicate the location of that individual's CE for the lottery. In a later study, Tversky and Kahneman (1992, pp. 305–306) used a two-stage list procedure to help them derive their probability weighting function. More recently, Holt and Laury (2002) adapted the list approach, keeping all payoffs constant while progressively changing the probabilities on both sides of the table to entail a single switching point intended to provide a measure of the individual's attitude to risk.

The advantage of the MPL method is that it is quite quick to administer and relatively straightforward to analyse as long as respondents do not switch from side to side more than once per table and as long as one assumes that the switching point corresponds with the SI point.[17] One disadvantage is that responses appear to be susceptible to various systematic framing effects connected to the ranges and frequencies selected by the researcher and the order in which rows are presented. Parducci (1965) provides an early discussion of range-frequency effects, Levy-Garboua et al. (2012) and Loomes and Pogrebna (2014a) report such effects in tables intended to elicit risk attitudes, and Sanchez Martinez et al. (2015) find evidence of such effects in the context of health state evaluation. Of course, if decisions are the result of some process of sampling and accumulation, it would not be surprising to find that "cues" provided by MPLs are liable to have various systematic effects. Further investigation of the magnitude of such effects may be required to make judgments about the appropriate trade-off between ease of administration and susceptibility to bias.

If we cannot ask each respondent to make a large number of choices and if the transparency of more compact iterative or list procedures makes them vulnerable to bias, a third approach may be considered for studies with large sample sizes. The dichotomous choice method, as advocated by Arrow et al. (1993) in a report for the National Oceanic and Atmospheric Administration, essentially involves identifying a set of possible values and then asking each of a very large number of respondents to make just one choice between the good under consideration and a single value picked at random from the set. On this basis, an aggregate "bid function" is derived, mapping from each value to the frequency with which the good is chosen by the subsample of respondents presented with that value.

To illustrate with reference to our first experiment, suppose that instead of asking 101 respondents to make 36 choices each—four choices at nine different levels of sure amount—we had been able to recruit 909 individuals and allocate them at random between each of the nine sure sums and ask each of them just once to choose between that sum and a particular bet. Since each person only sees a single sure amount and is unaware of the other amounts in the set, range-frequency effects are not in play; and since he makes a choice just once, he has the opportunity to deliberate carefully and does not need to be given filler tasks to distract him from previous choices. We might thereby arrive at data looking similar to Table 1, showing the frequency with which the bet is chosen at each level of sure sum.

Although we cannot infer any particular individual's SI from such data, we can at least identify the sample median and derive an estimate of the sample mean. Without knowing more about the nature of the underlying probability distributions from which each individual's response is drawn, it may be hard to say just how well the sample median or mean proxies the median or mean of individuals' SIs, but if we concur with the evidence from this and previous studies that WST broadly holds and if we have sufficiently large

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

samples making dichotomous choices across a large enough range of values, we may consider such estimates to be adequate for social cost–benefit purposes. Against that, the required sample sizes may make such data expensive to acquire; and under some conditions, even single questions may be subject to effects that distort the aggregate picture.[18]

In short, it seems unlikely that there is any one method that is entirely immune to some kind of effect or influence from the way it is framed or presented. Indeed, such effects may be unavoidable corollaries of the cognitive processes that generate decisions and cause them to take a probabilistic rather than a deterministic form. Our study cannot itself resolve the issue of which method to apply under any particular circumstances. However, our results draw attention to the importance of trying to develop descriptive models that not only accommodate probabilistic preferences but also provide a better understanding of the ways in which different elicitation procedures interact with underlying preferences.

Acknowledgments

The authors are grateful to the attendees at the Network for Integrated Behavioural Science 2015 Conference in Nottingham for useful comments and suggestions.

Endnotes

[1] Most often, this value has been elicited as a "reservation selling price"—that is, the individual is given ownership of the bet and is asked to state the smallest sure amount for which the individual would be prepared to sell the right to play out the bet. Sometimes the task is framed in terms of the maximum sure sum the individual would be prepared to pay to buy the right to play the bet and be paid accordingly. Lichtenstein and Slovic (1971) report experiments using both selling values and buying values. Luce (2000, Chapters 6 and 7) provides a detailed theoretical discussion of buying and selling prices for risky lotteries, together with some observations about the practical difficulties of eliciting such values.

[2] An intuitive explanation of how regret theory works in this context is as follows. An individual who behaves according to regret theory gives disproportionate weight to larger payoff differences within binary comparisons and is especially averse to being on the downside of such differences. In a $\{\$, M\}$ pair, the bigger difference is between the \$-bet's high payoff and the relatively modest sure sum offered by M, and this works against M and favours choosing \$. By contrast, in the $\{P, M\}$ pair, the higher payoff offered by P is typically not much better than the sure sum, whereas the more influential difference is between the sure sum and P's lower payoff, which works against P and favours M. Hence we may see $\$ \succ M$ and $M \succ P$ even when $P \succ \$$ in a direct comparison between those bets.

[3] Bostic et al. (1990, p. 204) expressed some concern that in their first experiment, the iterative procedure may have become so transparent that it affected respondents' answers by putting them into a valuation frame of mind. Experimental economists might also fear that a transparent procedure could encourage strategic answers intended to influence the options offered subsequently. The second procedure used by Bostic et al. (1990) was less transparent and therefore less vulnerable to those concerns.

[4] Instead of asking for a sure amount of money that the individual considers exactly as desirable as a particular bet, these questions ask the individual to state the probability of an even higher payoff that is regarded as equivalent to a particular bet. For example, if the P-bet were a 0.7 chance of £15 and a 0.3 chance of 0, an individual might be asked what probability p of £60 (with a 1-p chance of 0) would be exactly as desirable as that P-bet; that individual might also be asked what probability q of £60 (otherwise 0) would be exactly as desirable as a \$-bet offering a 0.3 chance of £40. Since p could be anywhere in the range between 0 and 0.7 without violating dominance, whereas q is constrained by dominance to lie between 0 and 0.3, there is more scope to give a higher probability equivalent for the P-bet while perhaps choosing the \$-bet in the direct choice between the two.

[5] A strong reversal is a case where the P-bet is chosen even though the stated certainty equivalent for the \$-bet is strictly higher than the maximum payoff offered by the P-bet. Thus, it amounts to an implicit violation of first-order stochastic dominance. Even though regret theory allows "standard" reversals, it cannot accommodate strong reversals.

[6] Here, we are abstracting from what have come to be called "trembles"—that is, cases where a transparently inferior option is chosen, perhaps because of some lapse of attention. For example, in cases where one option transparently dominates the other, it has been observed that the dominated alternative is chosen in a small proportion of cases—typically 1%–2%. Our figures, here and later, omit such trembles.

[7] When respondents are presented with a large number of tasks, there is necessarily a judgment to be made about the balance between task load and the quality of the data generated. However, it is not possible to conduct research into within-person variability without asking several series of questions, each repeated at least several times. Inevitably, this is liable to weaken the incentive per question. One of us, Loomes (2014), has cast doubt upon the quality of the data in a study by Guo and Regenwetter (2014), which asked respondents to make 1,600 choices between pairs of lotteries in approximately 80 minutes. However, the doubts about those data are based on an analysis of evidence of low sensitivity to parameter variations and abnormally high rates of violations of transparent dominance. Our task load was much lower than that in Guo and Regenwetter (2014), and as we shall see, our data exhibited good sensitivity to parameter variations and respectable levels of within-task coherence.

[8] The SI value is not well defined if $B = 0$ or if $B = 36$: in the former case, we can only conclude that SI $\leq$ £3.50 and in the latter case that SI $\geq$ 12.50; but for all other cases, each extra observation of the bet being chosen is counted as increasing the SI by £0.25.

[9] A good selection is listed at http://dfexperience.unibas.ch/literature.html (accessed July 5, 2015).

[10] http://gvp4c7.experimentaltoolbox.com/.

[11] Access to the experiment can be obtained by contacting the authors.

[12] This experiment was, in fact, one of two being conducted in parallel using the same general format of displays but with quite different questions. A total of 101 participants saw the questions relating to preference reversal as reported in this paper. Another 148 were presented with questions that were investigating the independence axiom—the results of which are reported in Loomes and Pogrebna (2014b). So in total, 249 individuals participated in one or other of the experiments, and the 100 who played a decision for real were drawn at random from the 249. At the time when individuals were deciding to take part, neither they nor we knew what the total number of participants would be, but the promise to pay 100 individuals seems to have been sufficiently attractive to induce a high (by DR@W standards) take-up.

[13] For both bets, estimated to the nearest £0.25.

[14] Throughout, we took the lower of the two numbers in the boxes to avoid any possibility that any overestimate could arise from taking the halfway point in the £0.10 difference.

Downloaded from informs.org by [155.246.103.35] on 05 April 2017, at 17:58 . For personal use only, all rights reserved.

[15] Of these, seven always stated the expected values of the bets, one always gave a value of £10, and eight always gave extreme values—seven stating the maximum payoff for each bet, with one stating the maximum payoff for the \$-bet and the minimum payoff for the *P*-bet.

[16] Prospect theory (Kahneman and Tversky 1979) and its more widely applied successor cumulative prospect theory (Tversky and Kahneman 1992) are exceptions in this respect. Prospect theory was formulated explicitly only as a theory of choice. When it came to deriving certainty equivalents as part of the process of estimating parameters for cumulative prospect theory, the authors were careful to arrive at those certainty equivalents via a choice-based procedure rather than by a direct elicitation (a point they emphasize in Tversky and Kahneman (1992, p. 306)). However, the structure of cumulative prospect theory entails the existence of certainty equivalents, and many subsequent authors have proceeded as if the model can be applied equally well to valuation/pricing tasks as to binary choice.

[17] In fact, it is common for a minority of respondents to switch sides more than once, usually leading to the exclusion of their data. Moreover, if underlying preferences are indeed probabilistic, the assumption that observed switching points will correspond with SI points may often be unsafe.

[18] For a more detailed discussion of variants of dichotomous choice methods compared with other valuation techniques used in surveys, and the various effects and disparities that have been reported, see Bateman et al. (2002), especially Sections 4.2.3 and 12.2.2.

References

Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H (1993) *Report of the NOAA Panel on Contingent Valuation* (Resources for the Future, Washington, DC).

Bateman I, Carson R, Day B, Hanemann M, Hanley N, Hett T, Jones-Lee M, et al. (2002) *Economic Valuation with Stated Preference Techniques: A Manual* (Edward Elgar, Cheltenham, UK).

Becker G, DeGroot M, Marschak J (1964) Measuring utility by a single-response sequential method. *Behavioral Sci.* 9(3):226–232.

Bell D (1982) Regret in decision making under uncertainty. *Oper. Res.* 30(5):961–981.

Blais A, Weber E (2006) A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment Decision Making* 1(1): 33–47.

Blavatskyy P (2009) Preference reversals and probabilistic decisions. *J. Risk Uncertainty* 39(3):237–250.

Blavatskyy P, Pogrebna G (2010) Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *J. Appl. Econometrics* 25(6):963–986.

Bostic R, Herrnstein R, Luce R (1990) The effect on the preference reversal phenomenon of using choice indifferences. *J. Econom. Behav. Organ.* 13(2):193–212.

Busemeyer J, Townsend J (1993) Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psych. Rev.* 100(3):432–459.

Butler DJ, Loomes GC (2007) Imprecision as an account of the preference reversal phenomenon. *Amer. Econom. Rev.* 97(1): 277–297.

Cohen M, Jaffray J-Y, Said T (1987) Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organ. Behav. Human Decision Processes* 39(1): 1–22.

Fishburn PC (1988) *Nonlinear Preference and Utility Theory* (Wheatsheaf Books, Brighton, UK).

Grether D, Plott C (1979) Economic theory of choice and the preference reversal phenomenon. *Amer. Econom. Rev.* 69(4):623–638.

Guo Y, Regenwetter M (2014) Quantitative tests of the perceived relative argument model: Comment on Loomes (2010). *Psych. Rev.* 121(4):696–705.

Holt C, Laury S (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.

Johnson J, Busemeyer J (2005) A dynamic, stochastic, computational model of preference reversal phenomena. *Psych. Rev.* 112(4): 841–861.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Levy-Garboua L, Maafi H, Masclet D, Terracol A (2012) Risk aversion and framing effects. *Experiment. Econom.* 15(1):128–144.

Lichtenstein S, Slovic P (1971) Reversals of preference between bids and choices in gambling decisions. *J. Experiment. Psych.* 89(1): 46–55.

Lindman H (1971) Inconsistent preferences among gambles. *J. Experiment. Psych.* 89(2):390–397.

Loomes G (2014) Quantitative tests of the perceived relative argument model: Reply to Guo and Regenwetter (2014). *Psych. Rev.* 121(4):706–710.

Loomes G, Pogrebna G (2014a) Measuring individual risk attitudes when preferences are imprecise. *Econom. J.* 124(576):569–593.

Loomes G, Pogrebna G (2014b) Testing for independence while allowing for probabilistic choice. *J. Risk Uncertainty* 49(3): 189–211.

Loomes G, Sugden R (1982) Regret theory: An alternative theory of rational choice under uncertainty. *Econom. J.* 92(368):805–824.

Loomes G, Sugden R (1983) A rationale for preference reversals. *Amer. Econom. Rev.* 73(3):428–432.

Loomes G, Starmer C, Sugden R (1991) Observing violations of transitivity by experimental methods. *Econometrica* 59(2): 425–439.

Luce RD (2000) *Utility of Gains and Losses Measurement Theoretical and Experimental Approaches* (Lawrence Erlbaum Associates, Mahwah, NJ).

Luce RD, Suppes P (1965) Preference, utility, and subjective probability. Luce RD, Bush RR, Galanter E, eds. *Handbook of Mathematical Psychology*, Vol. 3 (John Wiley & Sons, New York), 249–441.

MacCrimmon K, Smith M (1986) Imprecise equivalences: Preference reversals in money and probability. Working paper, University of British Columbia, Vancouver, BC, Canada.

Mosteller F, Nogee P (1951) An experimental measurement of utility. *J. Political Econom.* 59:371–404.

Otter T, Johnson J, Rieskamp J, Allenby GM, Brazell JD, Diedrich A, Hutchinson JW, MacEachern S, Ruan S, Townsend J (2008) Sequential sampling models of choice: Some recent advances. *Marketing Lett.* 19(3):255–267.

Parducci A (1965) Category judgment: A range-frequency model. *Psych. Rev.* 72(6):407–418.

Rieskamp J, Busemeyer J, Mellers B (2006) Extending the bounds of rationality: Evidence and theories of preferential choice. *J. Econom. Literature* 44(3):631–661.

Sanchez Martinez F, Pinto Prades J, Abellan Perpinan J, Martinez Perez J (2015) Avoiding preference reversals with opaque methods. Working paper, Glasgow Caledonian University, Glasgow, UK.

Seidl C (2002) Preference reversal. *J. Econom. Surveys* 16(5): 621–655.

Sopher B, Gigliotti G (1993) Intransitive cycles: Rational choice or random error? An answer based on estimation of error rates with experimental data. *Theory Decision* 35(3):311–336.

Stott H (2006) Cumulative prospect theory's functional menagerie. *J. Risk Uncertainty* 32(2):101–130.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4): 297–323.

Tversky A, Slovic P, Kahneman D (1990) The causes of preference reversal. *Amer. Econom. Rev.* 80(1):204–217.