ORIGINAL PAPER

Beyond choice: investigating the sensitivity and validity of measures of strength of preference

**David Butler · Andrea Isoni · Graham Loomes ·
Kei Tsutsui**

Abstract Many experiments investigating different decision theories have relied heavily on pairwise choices between lotteries. These are easy to incentivise, but often yield only limited dichotomous information. This paper considers whether respondents' judgments about their *strength of preference* (SoP) for one alternative over another can usefully supplement standard choice data. We report extensive evidence that such judgments show sensitivity to variations in question format and parameter values in the directions we should expect, not only within-subject but also between-sample. We illustrate how such judgments can usefully supplement standard pairwise choice data and enrich our understanding of observed behaviour.

D. Butler
School of Management and Governance, Murdoch University, Perth, Australia
e-mail: d.butler@murdoch.edu.au

A. Isoni (✉) · G. Loomes
Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, UK
e-mail: a.isoni@warwick.ac.uk

G. Loomes
e-mail: g.loomes@warwick.ac.uk

K. Tsutsui
Frankfurt School of Finance and Management, Frankfurt am Main, Germany
e-mail: k.tsutsui@fs.de

🙋 Springer

1 Introduction

During the course of the last five decades, data from laboratory experiments have challenged the descriptive validity of Expected Utility Theory (EUT) and have stimulated the development of numerous alternative models, which in their turn have been subjected to further experimental examination in the (so far unfinished) search for an adequate descriptive model (see Starmer 2000).

Much of this work has used pairwise choices between lotteries. This is natural enough, since most theories of choice under risk are formulated in terms of choices over primitives that take the form of lotteries over consequences, and lotteries with monetary payoffs can easily be produced under incentive-compatible laboratory conditions. However, while dichotomous choices show which alternative an individual chose, they do not reveal much more about the relative evaluation of the alternatives. Moreover, if individuals are heterogeneous in their preferences, any given set of predetermined alternatives may only discriminate between competing theories for a (possibly small) subset of the sample. As a consequence, choice-only experiments may only give an incomplete picture of the effects of interest.

In this paper, we consider a possible way to supplement the information provided by choice data. We propose a method that not only elicits incentive-compatible choices, but also asks participants to provide judgments about the *strength of preference* (SoP) for the alternative they choose. Such judgments are meant to provide some indication of the *relative degree of difference* between the two options as perceived by the decision maker. What is of interest is *not* the absolute value of SoP reported by the respondent but rather how SoP varies for the same individual across related decision problems. In the next section we expand on the theoretical status of the SoP concept and how it might add to our understanding of observed behaviour.

Our purpose here is to investigate the performance of an instrument which is simple to use and which might—perhaps after some further refinement—contribute to a large-scale comparison of various alternative theories in future experimental studies. However, since this instrument cannot be incentivised in the usual way, the credentials of this method need to be established empirically.

Thus our first objective is to investigate the sensitivity and validity of the new instrument. We do this by looking at how the self-reported SoP measures respond to unambiguous improvements in one of the options in a pair. These tests are particularly robust, as the kind of broad responsiveness we look for is predicted by all main competing theories. To preview our results, we find that the SoP judgments we elicit are very sensitive to *within*-subject manipulations and, somewhat to our surprise, even show sensitivity to *between*-sample manipulations.

Our second objective is to examine the usefulness of these SoP data. As examples of how the instrument might be deployed, we shall consider what additional insights it might provide into some issues of interest to decision theory: the stochastic element in people's preference; the robustness—or otherwise—of the notion of simple scalability in choice behaviour; and the status of the independence axiom of EUT.

The paper is organised as follows. Section 2 considers in more detail the concept of SoP and some possible implications for observable behaviour. Section 3 describes the instrument we used to elicit some measure of SoP. Sections 4 to 6 present our

Fig. 1 Mosteller and Nogee's
Fig. 2
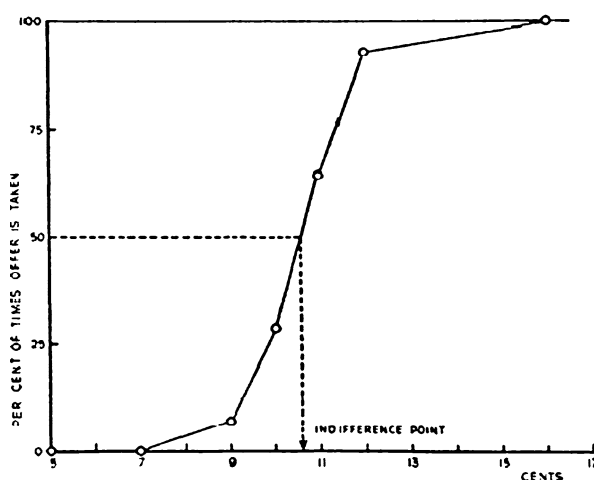
extensive sensitivity tests, illustrating some of the extra insights that SoP judgments can offer. Section 7 discusses the main limitations of our SoP instrument in its current form. Section 8 concludes.

2 Strength of preference: theoretical background

In their review of deterministic and probabilistic theories as they existed at that time, Luce and Suppes (1965) wrote: "...the intuitive idea of representing the strength of preference in terms of a numerical utility function—in terms of a subjective scale—is much too appealing to be abruptly dropped. Indeed, no-one has yet dropped it at all; every theory that we examine includes such a notion..." (pp. 332–333).

Early attempts to estimate individuals' utility functions—see, for example, Mosteller and Nogee (1951)—produced evidence of the probabilistic nature of those preferences. They found that if they made a gamble progressively better, the typical individual did not switch at some point from refusing the gamble 100 % of the time to accepting the gamble 100 % of the time—as would be supposed by deterministic theories with a single indifference point—but rather became increasingly likely (but not certain) to take the gamble over some range. Figure 1 reproduces the example Mosteller and Nogee reported of one individual who was repeatedly offered a gamble giving a 2/3 chance of winning X cents and a 1/3 chance of losing 5c, with X varied from one question to another. In the case in Fig. 1, the values of X were varied between 7c and 16c inclusive, with each value offered on 14 separate occasions. When $X = 7c$, the participant never accepted the bet. When $X = 16c$, he always took the bet. For values of X in between, the individual sometimes accepted the bet and sometimes rejected it, with the probability of accepting the bet increasing as X increased.

On this basis, Mosteller and Nogee estimated the value of X which would have led to the bet being accepted 50 % of the time and rejected 50 % of the time. As Fig. 1 shows, this rather primitive 'fitting' yielded an estimate of 10.6. However, for values

🖄 Springer

of X between 9c and 12c inclusive, there was some chance that the individual would accept the gamble on some occasions and reject exactly the same gamble on others.

One way of accounting for such behaviour is by supposing that on any particular occasion when he is asked whether to take the gamble or turn it down, the individual consults his preferences *as he perceives them at that moment*, with that perception consisting of some 'central tendency' difference between the subjective values of the alternatives, plus some variability of the kind that seems to be inherent to human judgment processes.[1]

For example, in the class of 'constant utility' models discussed by Luce and Suppes (1965, Sect. 5.2), the central tendency difference can be thought of as the difference between the EUs of the alternative options being considered.[2] However, the 'noisiness' of human judgment means that sometimes the direction of this difference is reversed.

With such a characterisation of the decision-maker's behaviour, Fig. 1 can be interpreted as follows. When $X = 10.6$, there is no difference between the individual's central tendency subjective values of rejecting or accepting the gamble. In this case, the actual choice is entirely decided by random noise in the judgmental process (modelled as an 'error term' with a median of 0). As the value of X becomes increasingly different from 10.6, the central tendency difference between the two options increases and thus is progressively less likely to be reversed by noise. So when X is 7c, the underlying difference favouring rejection is large enough that it is not reversed on any of the 14 presentations. And when $X = 16c$, the difference favouring acceptance is large enough that it is not reversed on any of the 14 presentations of the gamble with that payoff.

By asking people not only to record their choice but also to report their SoP for the option they are choosing, we are asking for an estimate of their perceived preferences at that moment, with that perception consisting of the central tendency difference between the subjective values of the options in conjunction with noise.

If an individual can make SoP judgments of this kind,[3] we might expect that sufficient repetition of the choice will allow us to cancel out the noise component and home in on a judgment about the *core* difference between the two alternatives. By doing this for a number of related pairs of alternatives, we can build up a picture of how core preferences are behaving, not only for the subset of individuals whose patterns of actual choices differ in the particular pairs considered, but also for all remaining

---

[1]Such patterns have been observed in many other judgmental contexts investigated by psychophysical experiments ever since Fechner (1860): for example, the probability of judging the lighter of two objects to be heavier increases as the true difference between their weights gets progressively smaller. Hence this class of models may be referred to as *Fechnerian*. There are other ways to make choices the outcome of a stochastic process—most notably the random preference approach (e.g. Becker et al. 1963; Loomes and Sugden 1995)—but for simplicity of exposition, we will restrict our attention to Fechnerian models, which have been particularly prominent in applied work.

[2]Although it seemed natural in 1965 to take EU as the 'core' deterministic theory, this is not the only possibility: the same broad idea can be applied to many of the numerous alternative core theories developed in the years since then.

[3]In fact, in decision analysis, it has been quite usual to elicit such judgments (see, for example, Von Winterfeldt and Edwards 1986).

 Springer

individuals whose final choices may be the same but who can be differentiated in ways that dichotomous choice alone would miss.

At the same time, we may be able to learn about the ways in which the noisiness of judgment varies with (or is independent of) the characteristics of the choices being made. Such knowledge may guide choices of stochastic specification when econometric methods are being used to try to compare the fit of different core theories. For example, Hey and Orme (1994) fitted a number of alternative core functions to their dataset on the assumption that the variance of the error term was constant across all pairs of lotteries in their experiment, whereas Buschena and Zilberman (2000) proposed a heterogeneous error structure. In Sect. 4, we shall begin to explore this issue in the context of pairs in which one lottery first-order stochastically dominates the other and where we can systematically vary each dimension of the differences between the alternatives.

Figure 1 reveals another interesting way in which SoP judgments can be used. In a Fechnerian model, the slope of the fitted line depends on the variance of the error term, the curve becoming flatter as the variance increases. Suppose the constant loss in each of the gambles offered by Mosteller and Nogee was increased to 7c. If we assumed that the error term has constant variance, we would expect the resulting curve to be just a parallel shift to the left, with an important implication: for any value of the win on the horizontal axis, the probability of taking the gamble with a loss of 7c would be strictly less than the probability of taking the gamble with a loss of 5c. This property is known as 'simple scalability', or 'independence principle in probabilistic choice' (see Tversky and Russo 1969). The property would not hold if, for instance, the variance of the error term was larger for the loss of 7c than for the loss of 5c. In that case, the curve for the loss of 7c would be flatter, allowing for the possibility that, for some values of the winning amount, the probability of taking the gamble with the larger loss would be greater than that of taking the gamble with the smaller loss.

More formally, let $\Pr(J, K) = f[v(J), v(K)]$ denote the probability that lottery J is chosen over lottery K as an increasing function of the subjective value/utility assigned to J, $v(J)$, and a decreasing function of the subjective value/utility of K, $v(K)$. Simple scalability implies that $\Pr(J, L) > \Pr(K, L)$ iff $\Pr(J, M) > \Pr(K, M)$ for all L and M: that is, order independence from the comparators L and M. Under the assumption of constant variance of the noise component, $\text{SoP}(J, K)$ can be treated as a proxy for $f[v(J), v(K)]$, allowing us to use individual-level data to check if simple scalability is satisfied. If simple scalability is not respected, our SoP judgments would indicate that the assumption of constant variance is inappropriate.[4] We present an application along these lines in Sect. 5.

Finally, SoP measures may give us greater information with which to discriminate between competing models. To illustrate, consider two lotteries which involve

---

[4]Systematic violations of simple scalability have been widely documented in the literature using sample proportions as proxies for choice probabilities of a representative agent (see Busemeyer and Townsend 1993 and references therein). With SoP judgments it is easier, with just a few observations, to form a picture of whether simple scalability is satisfied *at the individual level*, which is harder to do using binary choices alone (for an early example, see Busemeyer 1985).

	1-65		66-100
A	£13		£7
	65%		35%

	1-40		41-100
B	£37		0
	40%		60%

Fig. 2 How the lotteries were displayed

some probability p of a payoff X that is common to both of them. For example, suppose lottery S offers a 0.5 chance of £12 and a 0.5 chance of 0—more compactly, S = (£12, 0.5; £0, 0.5)—while lottery R = (£25, 0.4; £0, 0.6), so that they have in common a 0.5 chance of 0. The core difference between them is $v(S) - v(R)$. EUT implies that if we create two new lotteries by replacing the 0.5 chance of 0 with a 0.5 chance of £25, to give S′ = (£25, 0.5; £12, 0.5) and R′ = (£25, 0.9; £0, 0.1), the difference is unaffected—i.e. $[v(S') - v(R')] = [v(S) - v(R)]$. However, a number of alternative theories to EUT entail $[v(S') - v(R')] > [v(S) - v(R)]$. If we rely solely on observing one-off choices, the only cases which we can use to discriminate between EUT and those alternative theories are cases where the difference changes from positive to negative or from negative to positive—that is, where an individual chooses S′ over R′ but chooses R over S, or else where an individual chooses R′ over S′ but chooses S over R. Both of these depart from EUT whereas the first but not the second is compatible with those alternative models. However, if the total number of these cases is small and/or the asymmetry is not very pronounced, it may be difficult to rule out the possibility that these observations are produced purely by the noise in people's judgments.

On the other hand, if we can elicit repeated SoP measures for each pair, we may be able to get less noisy estimates of the core differences and compare those estimates of the differences $[v(S') - v(R')]$ and $[v(S) - v(R)]$ for all respondents and not just for the subset of cases where the sign changes. We illustrate how SoP judgments can be employed in comparisons of this kind in Sect. 6.

3 Our strength of preference instrument

The properties of our SoP instrument were investigated in an experiment run in 12 sessions at the University of East Anglia in May 2009. The sample consisted of 138 individuals who were randomised between two subsamples (which we label V and W). Each participant made 100 choices between pairs of lotteries on the understanding that when they had completed all 100, one would be picked at random and played out for real, with the entire payment for their participation depending on how their decision in that one question played out (i.e. there was no 'show-up' fee or any other source of monetary reward). The experiment took less than an hour to complete, and participants earned an average of £10.25.

The pairs of lotteries were displayed using the format illustrated in Fig. 2. It was explained that the outcome for any lottery would be determined by the respondent
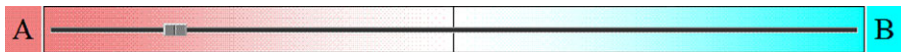
Fig. 3 The 'strength of preference' instrument

drawing at random a disc from an opaque bag containing 100 discs each with a different number inscribed upon it.[5] So, for example, if lottery A in Fig. 2 were being played out, the individual would receive £13 if the randomly-drawn disc bore a number between 1 and 65 inclusive (as displayed above the sum) or would receive £7 if the disc bore a number in the 66–100 range. The chances of winning any sum of money were given in percentage form underneath each sum—in this case, 65 % and 35 % respectively.

For each pair, the respondent was asked not only to choose the alternative he preferred, but also to use the bar shown in Fig. 3 to record his perceived SoP for the chosen alternative.

The instrument worked as follows. When a fresh pair of lotteries was first presented, the button sat at the centre of the bar. It was moved by clicking on it and dragging it either to the left to signify a preference for the lottery labelled with a red A, or to the right to signify a preference for the lottery labelled with a blue B. The respondent was told: "If you feel that both alternatives are almost equally good so that you think the one you are choosing is only SLIGHTLY better than the other one, just move the button a little way in the direction of your choice. However, if you think the one you are choosing is VERY MUCH BETTER than the other one, move the button a long way along the bar in the direction of your choice, possibly as far as the end if you feel very strongly indeed. Once you have moved the button to the position that shows which alternative you choose and how much better you think it is, press OK. Then you will be asked to confirm your choice (or change it, if you change your mind) before moving to the next decision."

In fact, as participants saw when they moved the cursor, some accompanying text appeared and changed with the position of the button. Although it was not apparent to participants, the bar on each side of the centre was calibrated to 100 points, corresponding to a score ranging from 0 to 200 from one end to the other. When the button was located at any of the first 25 points to each side of the centre, text appeared which read 'You think A(B) is SLIGHTLY better'. When the button was moved to anywhere on the next 25 % of the bar, the text changed to 'You think A(B) is BETTER'. For the next quarter of the bar the text read 'You think A(B) is MUCH better'. And for the quarter furthest from the centre, it read 'You think A(B) is VERY MUCH better'. When they confirmed their decision, the exact position of the cursor on the 200-point scale was recorded.

Because of this feature of the instrument, we anticipated that some participants might use the slider as an 8-point scale, with four levels on each side. Figure 4 shows a histogram of the SoP values recorded by the 138 participants in the 100 tasks (a total of 13,800 decisions), normalised over a 100-point scale to represent the SoP in favour of the chosen lottery in a pair.

---

[5] The full text of the instructions is reproduced in the electronic supplementary material (ESM).

Ⓐ Springer

Fig. 4 Aggregate distribution
of normalised SoP score
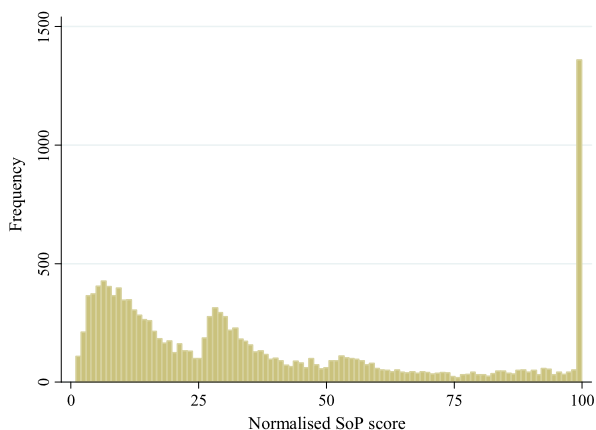
Those who are concerned that the lack of direct incentives might result in participants using the slider in a minimalist or random way should be reassured by the evidence that the whole scale was used. On the other hand, there are clear spikes just after 25 and 50, suggesting that many participants were sensitive to the verbal descriptions and more often ceased moving the slider soon after a change of wording had been triggered.[6] In the next three sections, we will take this aspect of the data into account by coding the SoP scores recorded in the experiment to map onto an 8-point scale. Where necessary, we will refer to the four levels of SoP on each half of the scale as 'vmb' (standing for 'very much better'), 'mb' ('much better'), 'b' ('better') and 'sb' ('slightly better').

When information about the individual *decision* is required, we simply assign the score recorded for that decision on the 200-point scale to the corresponding category on the 8-point scale. However, as explained above, repetition of the SoP judgments can give a more accurate picture of the central tendency difference between the subjective values of the two lotteries involved. For this reason, the great majority of lottery pairs were faced by participants at least twice. Whenever information about a particular *pair* is required, in order to determine which of the eight categories each participant's SoP for that pair falls into, we first compute the average of the SoP score recorded on the original 200-point scale for the various occasions in which the pair was faced by the participant. We then assign the resulting average to the corresponding category. This gives a less noisy representation of the participant's preferences, and ensures the independence of observations across participants for our statistical tests.[7] However, it should be born in mind that the distributions of SoP categories that we present are to be interpreted as reflections of the participants' *estimated balance of preference* for one alternative over the other, even though they may sometimes have chosen differently on different occasions. For simplicity, we will sometimes find it

---

[6]This suggests that people's ability to discriminate their SoP level may be confined to a limited number of categories, a tendency that has been found in many other domains since Miller (1956).

[7]This procedure results in a missing value whenever the average falls exactly at the midpoint of the scale, where the slider was started. However, this happens very rarely (in fewer than 0.7 % of the cases).

🙋 Springer

convenient to refer to the distributions as if they were choices, but what we really mean is the preference implied by averaging the SoP responses.

There is also a very prominent spike at 100 that occurs with a frequency of 1359. However, 883 (almost two thirds) of the cases in which maximal SoP was recorded correspond to pairs in which one of the lotteries first-order stochastically dominates the other, suggesting that some participants may have used the slider not simply to indicate their SoP but also to express their *confidence* that they made the right decision. The relationship between SoP and confidence will be considered in the discussion.

We do not suppose that all individuals interpret the wording of these SoP tasks in the same way. As we shall see, the data suggest that different individuals show different sensitivity when using the SoP scale. We suppose only that any particular individual will display a degree of internal consistency in the way he/she interprets and applies the SoP scale.

With our instrument, SoP judgments are recorded on a bounded scale. We took this option because we felt that allowing for SoP to be reported on an open-ended scale would not have bought us much extra: for the range of payoffs in most experiments, a simple bounded scale would appear adequate. More sophisticated arguments could be adduced—for example, a more elaborate way of justifying boundedness would be to introduce some kind of scaling of the error term along the lines proposed by Blavatskyy (2011) or Wilcox (2011).[8] We hope it will become evident that the bounded nature of the scale does not represent a serious limitation of the instrument.

4 The first-order stochastic dominance pairs and variability in SoP judgments

In this Section, we focus on 20 of the pairs presented, in which one of the two lotteries first-order stochastically dominates the other. We use these 'FOSD' pairs for two main goals. The first is to explore to what extent the SoP instrument responds to unambiguous improvements in one of the options when, arguably, there is a correct answer. The second is to illustrate how SoP judgments can shed light on the characteristics of the noise component in people's preferences.

Table 1 summarises the design. There were two 'baseline' lotteries, as shown in the left-hand column: one offered a $p = 15\%$ chance of receiving payoff $b = £15$ and an 85 % chance of 0, while the other offered a $p = 35\%$ chance of $b = £35$ and a 65 % chance of 0. These were paired with alternatives that dominated them in some way: either by holding the probability of winning at p but increasing the payoff by an increment m, with m being either £1 or £10; or else holding the positive payoff at b but increasing the probability of winning it by an increment q, with q being either 1 % or 10 %. This part of the design was primarily intended to see whether the SoP instrument was sensitive to different values of m and/or q, given the same baseline.

We were also interested to see whether the immediacy/transparency of the dominance relationship had an impact on subjects' responses. For half of the displays, the

---

[8]A bounded SoP can also be achieved in the random preference framework by, for example, assuming that preferences are represented by a family of utility functions with risk aversion coefficients drawn from a closed interval.

Table 1 Summary of the FOSD pairs

Baseline	Increment			Repetitions	
	m (£)	q (%)	O/D	sub. V	sub. W
b = £15, p = 15 %	1	0	O	2	1
	1	0	D	2	
	0	1	O	2	1
	0	1	D	2	
	10	0	O	2	1
	10	0	D	2	
	0	10	O	2	1
	0	10	D	2	
b = £35, p = 35 %	1	0	O	1	2
	1	0	D		2
	0	1	O	1	2
	0	1	D		2
	10	0	O	1	2
	10	0	D		2
	0	10	O	1	2
	0	10	D		2

O = Overlapping display;
D = Disjoint display

two states offering the positive payoffs were lined up so that respondents could see either that every numbered disc which would pay b in the baseline lottery would also pay b + m in the dominating lottery, or else that every numbered disc which would pay b in the baseline lottery would also pay b in the dominating lottery but that another q adjacent numbered discs would also pay b in the dominating lottery rather than 0 in the dominated lottery: we refer to displays in this form as *overlapping*, indicated by an O in Table 1. For the other half of the FOSD pairs, the positive payoffs were at opposite ends of the displays, so that they required more care and attention in order to identify the stochastic dominance relationship: we refer to these displays as *disjoint*—D in Table 1.

Finally, in the majority of cases each pair was presented twice, with the A-B ordering reversed: this is indicated by a 2 in one of the last two columns of Table 1, and was intended to control for any effect the A-B ordering might have on responses.[9] Repeating the tasks also allows us to tap into the intrinsic variability that characterises reported SoP judgments. There were some instances where a particular pair was presented to members of one subsample just once in a particular format that was identical to one of the pairs presented to the other subsample, in order to allow a direct check that both subsamples were answering those questions in a way that was

[9]These 20 FOSD pairs were interspersed among the 100 pairs. We ensured that each of these pairs was faced once before any of them was presented for the second time. The sequence of pairs was predetermined, but differed for the two subsamples.

🖄 Springer

not systematically different:[10] these cases are indicated by a 1 in the last two columns of Table 1.

The results of the sensitivity tests are summarised in Table 2, which reports, separately for the two subsamples, the distributions of responses classified according to the different SoP categories. The data are arranged so that +vmb, +mb, +b and +sb reflect decreasing preference for the domina*ting* alternative. Similarly, −sb, −b, −mb and −vmb reflect increasing preference for the domina*ted* alternative. Each row refers to one of the pairs that were presented twice. For example, the row with **Base = 15, m = 1, q = 0, D**—that is, the second row for Subsample V—shows the distribution of average SoP judgments for the pair in which the baseline (dominated) option offered a 15 % chance of £15 while the dominanting alternative offered a 15 % chance of £16, and where the display was in disjoint format, so that the positive payoffs were at opposite ends of the display. To give another example, the row with **Base = 35, m = 0, q = 10, O**—the penultimate row for Subsample W—shows the distribution of SoP judgments for the pair in which the baseline was a 35 % chance of £35 while the dominanting alternative was a 45 % chance of £35, and where the display was in overlapping format, so that the positive payoffs were at the same end of the display.

When dominance is easiest to spot—i.e. in the overlapping displays—there are only a small minority of violations: for the eight cases, four from each subsample, when there are overlapping displays, there are 8 responses which violate dominance, constituting roughly 1.4 % of the observations. When dominance is less immediately transparent—in the disjoint displays—the number of violations rises to 55, constituting roughly 10 % of the observations.[11]

The sensitivity of SoP responses to the magnitude of the difference, either on the payoff dimension or on the probability dimension, is strongly indicated by the aggregate patterns in Table 2, but Table 3 presents the relevant within-person data. Each column shows the number of categories between the SoP registered by an individual in the case involving the greater m or q and the SoP registered by that same individual in the corresponding case involving the smaller m or q. For example, the row labelled **15, q = 25 − 16, O** shows how each individual's SoP for a 25 % chance of £15 over a baseline 15 % chance of £15 compared with that same individual's SoP for a 16 % chance of £15 over the same baseline, when the displays in both cases were overlapping. Absent violations of dominance, the maximum number of categories between the two SoP values should be 3 (if the dominating lottery is rated vmb than the dominated lottery for the larger increment and is rated sb for the smaller increment), with values between +1 and +3 being associated with the expected responsiveness of the SoP measures to the size of the increment in money or probability. Values greater than 3 mean that dominance has been violated in the pair with the smaller increment.

---

[10] The evidence—reported in more detail in the ESM—suggests that the random assignment of participants to subsamples V and W was effective in ensuring that there were no systematic differences between the two subsamples.

[11] Due to the way the data in Table 2 are coded (by assigning the average SoP score over the two repetitions to the corresponding SoP category), these numbers differ from the raw count of separate violations. In O pairs, we observed a total of 27 violations (out of a total of 1104 observations). In D pairs, the corresponding number was 133.

Table 2 SoP Distributions in FOSD pairs

Task				SoP categories based on mean scores								SoP range		Dom. Viol.
Base	m	q	O/D	+vmb	+mb	+b	+sb	−sb	−b	−mb	−vmb	Mean	Std. dev.	
Subsample V														
15	1	0	O	12	6	10	39	1	0	0	0	0.54	1.35	1
15	1	0	D	10	5	6	40	5	3	0	0	0.64	1.10	8
15	0	1	O	8	6	11	42	0	0	1	0	0.65	1.26	1
15	0	1	D	4	6	3	42	8	2	2	1	0.96	1.49	13
15	10	0	O	26	10	17	15	0	0	0	0	0.64	1.28	0
15	10	0	D	16	11	15	21	3	1	0	0	1.29	1.90	4
15	0	10	O	17	14	20	16	1	0	1	0	1.10	1.16	2
15	0	10	D	13	10	14	21	7	1	0	1	1.13	1.54	9
Subsample W														
35	1	0	O	16	8	12	32	1	0	0	0	0.45	0.76	1
35	1	0	D	11	5	14	32	3	2	0	0	0.84	1.36	5
35	0	1	O	14	10	14	30	1	0	0	0	0.68	1.10	1
35	0	1	D	6	4	11	38	6	2	0	0	1.16	1.76	8
35	10	0	O	22	16	23	7	0	1	0	0	0.61	0.97	1
35	10	0	D	16	15	25	8	4	0	0	0	1.16	1.59	4
35	0	10	O	18	21	17	12	0	1	0	0	0.91	1.01	1
35	0	10	D	15	13	25	12	2	1	1	0	0.87	1.11	4

Table 3 SoP sensitivity in FOSD pairs

Comparison	SoP difference (cat. based on average)						Sig.[a]
	<0	0	1	2	3	>3	
Subsample V							
$15, m = 25 - 16,$ O	1	32	22	2	11	0	***
$15, m = 25 - 16,$ D	6	26	22	5	6	2	***
$15, q = 25 - 16,$ O	4	29	18	14	3	0	***
$15, q = 25 - 16,$ D	7	25	21	6	3	4	***
Subsample W							
$35, m = 45 - 36,$ O	4	30	24	9	2	0	***
$35, m = 45 - 36,$ D	4	25	21	13	2	1	***
$35, q = 45 - 36,$ O	7	32	18	10	2	0	***
$35, q = 45 - 36,$ D	6	19	24	15	3	0	***

[a]Significance in a one-tail Wilcoxon test on category based on average SoP: *10 %, ** = 5 %, *** = 1 %

Negative numbers do not necessarily reflect violations of dominance, but may be associated with cases in which, contrary to our expectation, a stronger preference is registered for the smaller increment. The last column of the table reports tests of the null hypothesis that the distributions of SoP are not different for the two pairs against the alternative that they are more skewed towards the dominating lottery in the pair with the larger increment.[12]

These data clearly show movements in the direction we should expect. In every comparison, except **35**, **q = 45 − 36**, **O**, the majority of the observations for which the SoP for the smaller and the SoP for the larger increment differ are in the range between +1 and +3, which is consistent with participants reporting a stronger preference for the dominating option when the increment is larger. All differences are significant in the predicted direction at the 1 % level. Moreover—and arguably, even more strikingly—the appropriate patterns of movement are not only evident *within* sample but also *between* samples. That is, if we splice together the responses from both subsamples we find that as the EV differences increase from £0.15 to £0.35 to £1.50 to £3.50, the numbers of +sb responses go steadily down, while the numbers of responses reporting greater levels of SoP systematically rise. This same trend is observable for all four combinations of money/probability increments and overlapping/disjoint displays. The fact that we find this pattern even when the difference is quite small (£0.35 vs £0.15) *and* when these two differences are being evaluated by different and independent subsamples, is a clear sign of the sensitivity of the SoP instrument in this context.

As we explained in Sect. 2, SoP judgments can shed light on the behaviour of the noise element that characterises people's preferences. The second and third to last columns of Table 2 report the mean and standard deviation of a statistic called SoP

---

[12]These Wilcoxon signed-rank tests are based on the distributions of SoP categories obtained using the average SoP in the two observations of each pair.

Ⓐ Springer

Table 4 Variability of SoP in FOSD pairs

	Money vs. Probability		Small vs. Large		Overlapping vs. Disjoint	
	Sub V	Sub W	Sub V	Sub W	Sub V	Sub W
FOSD pair set 1	*Money*		*Small*		*Overlapping*	
Mean range	0.78	0.76	0.70	0.78	0.73	0.66
Std. dev. range	0.92	0.62	0.81	0.80	0.86	0.59
FOSD pair set 2	*Probability*		*Large*		*Disjoint*	
Mean range	0.96	0.91	1.04	0.89	1.00	1.01
Std. dev. range	0.84	0.68	0.93	0.71	1.00	0.80
Sig.[a]	*	*	***		**	***

[a]Significance in a two-tail Wilcoxon signed-rank test: * = 10 %, ** = 5 %, *** = 1 %

range. For each of the FOSD pairs which were repeated twice, this statistic is obtained by taking the sample average of the absolute difference between the SoP categories recorded in the two instances each pair was faced by each participant. If the assumption of constant variance is appropriate, we should see no difference in this statistic between cases in which the increment was small and those in which it was large, no difference between cases in which the increment was on the money rather than on the probability dimension, and no difference between overlapping and disjoint displays. The data in Table 2 show a different picture. Within each subsample, SoP judgments are more variable for cases involving the larger increment than for the corresponding cases involving the smaller increment in all eight possible comparisons; they are more variable when dominance is obtained by altering the probability rather than the money dimension in seven out of eight cases; and are also more variable for overlapping than for disjoint displays in seven out of eight cases.

Table 4 reports some relevant statistical tests. For each of the three comparisons (money vs. probability, small vs. large and overlapping vs. disjoint), we compute the average of the SoP range for the four pairs corresponding to each level of the relevant dimension (e.g. four money pairs and four probability pairs), separately for each sample. We then test the hypotheses that the corresponding average variables do not differ between the two levels of that dimension. We reject the null hypothesis in five out of six cases (twice at the 1 %, once at the 5 %, and twice at the 10 % level), in the direction that one would expect. There are, however, some qualitative differences between the two subsamples: the greatest contributor to SoP variability is the increment size in Subsample V (an average difference of 0.34 SoP categories), while the display format has greatest influence in Subsample W (also a difference of 0.34 categories on average).

Thus our SoP judgments reveal that a model in which noise is assumed to have constant variance across lottery pairs may be erroneous, and suggest ways in which variance should be modelled to achieve a better description of the data.

🖄 Springer

5 The preference reversal pairs and simple scalability

52 of the pairs faced by participants were built around lotteries with the structure of those commonly used in studies of the Preference Reversal (PR) phenomenon. The classic preference reversal phenomenon (see Lichtenstein and Slovic 1971; and for a survey, Seidl 2002) revolves around two lotteries: one—the $-bet[13]—offers a relatively high payoff with a fairly small probability of receiving it; while the other—the P-bet—offers a much greater probability of a more modest payoff. A 'standard' reversal occurs when individuals place a higher certainty equivalent value on the $-bet than on the P-bet but prefer the P-bet over the $-bet when asked to make a straight choice between the two. The opposite reversal—placing a higher value on the P-bet but picking the $-bet in a straight choice—is relatively rarely observed. A different asymmetry has also been reported, although fewer studies have looked for it. This involves individuals being asked to report their probability equivalents for each bet—i.e. the smaller probability of some payoff higher than the payoff of the $-bet that they regard as exactly as good as a given bet—as well as making a straight choice between the two: in these tasks, it is more common to observe someone who chooses the $-bet place a higher probability equivalent on the P-bet than choose the P-bet and place a higher probability equivalent on the $-bet (see, for example, Butler and Loomes 2007).

These 52 choices were made up as follows. On four different occasions respondents were asked to choose between a $-bet and a P-bet. For both subsamples, the $-bet was always the same: $ = (£40, 0.25; £0, 0.75). For subsample V, the P-bet was P1 = (£10, 0.9; £0, 0.1), while for subsample W, it was the unambiguously less attractive P2 = (£10, 0.65; £0, 0.35). We also asked each respondent to choose between each bet and four levels of sure amount (£10, £8, £6 and £4) and four prospects offering different chances of winning £60, which we denote by R0.25 = (£60, 0.25; £0, 0.75), R0.20 = (£60, 0.2; £0, 0.8), R0.15 = (£60, 0.15; £0, 0.85) and R0.10 = (£60, 0.1; £0, 0.9) and which we shall refer to in this context as the R-bets. Each of those choices was presented to respondents on three different occasions within the session. Thus, in total, each respondent was asked to make: 4 choices between $ and P; 12 choices between $ and four levels of sure amount; 12 choices between the relevant P-bet and four levels of sure amount; 12 choices between the $-bet and the four R-bets; and 12 choices between the relevant P-bet and the four R-bets.[14]

This design allows us to see how the SoP instrument responds to changes in one prospect while the other is held constant. For instance, one should expect that as the certainty is progressively reduced in £2 steps from £10 to £4, the SoP for a particular $- or P-bet should progressively increase. Since we used two different P-bets in the two subsamples, our design also allows for between-subject comparisons.[15]

The data are reported in Table 5, using the convention that the first lottery of each pair {S, R} is relatively *safe* (S), while the second is relatively *risky* (R). For all cases in which S (R) is chosen, the number of instances in which the average SoP results in

---

[13]From now on, we will use the words 'lottery', 'bet' and 'prospect' interchangeably.

[14]As with the FOSD pairs, these pairs were interspersed among the 100 pairs and their order was predetermined and different for the two subsamples. We ensured that each pair was faced once before it was

Table 5 SoP distributions in PR pairs

Pair {S, R}	S preferred				R preferred				S (%)
	vmb	mb	b	sb	sb	b	mb	vmb	
{£10, P2}	55	7	6	0	0	1	0	0	99 %
{£8, P2}	11	18	25	12	0	2	1	0	96 %
{£6, P2}	4	10	15	29	9	1	0	0	85 %
{£4, P2}	3	1	7	24	27	6	1	0	51 %
{£10, P1}	45	11	8	4	1	0	0	0	99 %
{£8, P1}	5	4	25	14	17	3	1	0	70 %
{£6, P1}	1	2	12	13	21	14	3	3	41 %
{£4, P1}	0	0	4	7	24	22	6	6	16 %
{£10, $}	8	16	45	47	15	4	0	2	85 %
{£8, $}	4	7	27	61	28	7	3	1	72 %
{£6, $}	2	5	24	59	35	9	1	2	66 %
{£4, $}	2	3	7	49	44	21	9	3	44 %
{P1, $}	2	4	15	31	11	4	0	2	75 %
{P2, $}	1	0	7	40	17	3	1	0	70 %
{P1, R0.10}	4	6	17	32	5	3	0	2	86 %
{P1, R0.15}	1	1	15	37	10	3	1	1	78 %
{P1, R0.20}	0	2	16	29	15	5	1	1	68 %
{P1, R0.25}	1	0	12	29	17	7	2	1	61 %
{P2, R0.10}	1	0	15	44	9	0	0	0	87 %
{P2, R0.15}	1	0	7	47	12	2	0	0	80 %
{P2, R0.20}	1	1	6	32	21	7	1	0	58 %
{P2, R0.25}	0	0	6	31	22	9	1	0	54 %
{$, R0.10}	4	8	45	72	8	0	0	1	93 %
{$, R0.15}	3	3	22	85	18	5	1	1	82 %
{$, R0.20}	1	0	5	49	50	20	11	2	40 %
{$, R0.25}	0	0	1	2	9	33	39	54	2 %

it being rated vmb, mb, b and sb than R (S) are reported. The final column of the table reports the percentage of the sample whose preference is in favour of the safer lottery in the pair. Note that the data for pairs involving the $-bet and the certainties or the

---

presented for the second time, and twice before it was presented for the third time. Which lottery was presented as A was kept constant in all repetitions.

[15]We checked for any systematic changes in the SoP distributions when tasks are repeated several times. A non-parametric test comparing the average SoP for the first presentation of all PR pairs with the average SoP for the last repetition finds no evidence of systematic trends over time.
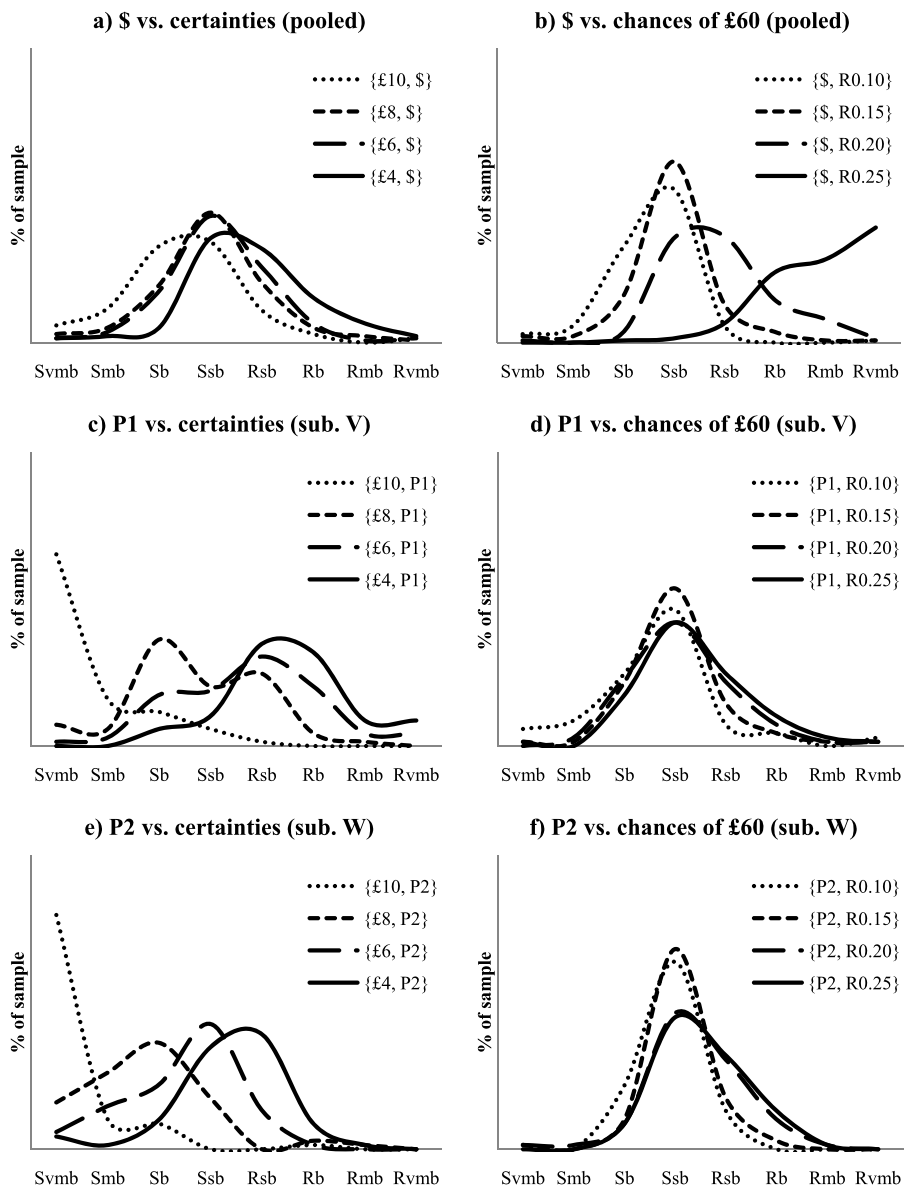
Springer

a) $ vs. certainties (pooled)

{£10, $}
{£8, $}
{£6, $}
{£4, $}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

b) $ vs. chances of £60 (pooled)

{$, R0.10}
{$, R0.15}
{$, R0.20}
{$, R0.25}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

c) P1 vs. certainties (sub. V)

{£10, P1}
{£8, P1}
{£6, P1}
{£4, P1}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

d) P1 vs. chances of £60 (sub. V)

{P1, R0.10}
{P1, R0.15}
{P1, R0.20}
{P1, R0.25}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

e) P2 vs. certainties (sub. W)

{£10, P2}
{£8, P2}
{£6, P2}
{£4, P2}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

f) P2 vs. chances of £60 (sub. W)

{P2, R0.10}
{P2, R0.15}
{P2, R0.20}
{P2, R0.25}

% of sample

Svmb Smb Sb Ssb Rsb Rb Rmb Rvmb

Fig. 5 SoP distributions in PR pairs

R-bets are pooled as the two subsamples reported SoP distributions which were not significantly different in any of the eight comparisons.[16]

The data are depicted in Fig. 5. For each of $, P1 and P2, the figure shows the SoP distributions (taken from Table 5) for choices between the lottery and an alternative

---

[16]Two-tailed Mann-Whitney tests: see ESM for details.

that is made progressively more attractive (moving from £4 to £10 in the left side of the figure, or from R0.10 to R0.25 in the right side). In order to ease comparisons between the various diagrams, the vertical axis of each graph reports relative frequencies obtained from the numbers in Table 5 by normalising with respect to the total number of observations (recall that the data for the $-bet are pooled across the two subsamples).

The SoP distributions show clear responsiveness to changes in parameters.[17] The trends are particularly strong for the comparisons involving the P-bets and certainties (panels c and e in Fig. 5), and for the pairs involving the $-bet and the R-bets (panel b). They are also apparent but perhaps less pronounced for the $-bet and the certainties (panel a) and evident—although less easy to see instantly—in the comparisons between the two P-bets and the R-bets (panels d and f). In general, and in line with expectations from earlier work (e.g., Butler and Loomes 2007), it seems that the distinctions become less sharp when the comparisons are between alternatives that are more dissimilar. So, for {P1, $} and {P2, $} (not shown in Fig. 5, but reported in Table 5), there is still a systematic trend, but it is less pronounced than in the comparisons with sure amounts. As the comparator becomes even more distanced from the P-bets, as in the case of the choices between P-bets and R-bets, the trend, while still discernible, becomes somewhat fuzzier. However, all differences between the SoP distributions involving the certainties and the R-bets, except for the comparisons between {P1, R0.20} and {P1, R0.15}, are strongly significant in the predicted direction in within-subject comparisons of the SoP distributions.[18]

There is also a substantial degree of *between*-subject responsiveness. Recall that we used two different P-bets in the two subsamples, P2 being dominated by lottery P1. When these lotteries are compared with a common alternative, it seems not unreasonable to expect that the common alternative will be preferred more strongly to P2 than to P1. Such patterns are clearly visible in panels c and e in Fig. 5. For each level of certainty (£X), the distribution for {£X, P1} lies to the left of that for {£X, P2}. A similar tendency, though not equally pronounced, can be observed in panels d and f for comparisons with the R-bets.[19]

The responsiveness we have just highlighted relates to the issue of simple scalability that we mentioned in Sect. 2. Previous studies reporting evidence of systematic violations of simple scalability using choice proportions have suggested that differing degrees of similarity are often implicated in such violations. And so it is in our data.

Since the P-bets are considerably more similar to certainties than is the $-bet, we see in Table 5 and Fig. 5 the SoP scores being much more responsive to the changes in the sure amounts when these are being compared with P-bets than when they are being compared with the $-bet. In the first four rows of Table 6, the subsample average SoP responses are reported for comparisons between £8 and each bet and be-

---

[17]The spikes at vmb in the distributions for pairs {£10, P1}, {£10, P2} and {$, R0.25} are due to the fact that these are pairs in which one of the two lotteries dominates the other. See Fig. 4 and Sect. 7.

[18]These comparisons are based on the Wilcoxon signed-rank test and use the SoP distribution obtained from averaging the score recorded in the three presentations of each task. See ESM for details.

[19]Five out of eight of these between-subject differences are large enough to reach statistical significance (see ESM).

Springer

Table 6 SoP and violations of simple scalability

	Sub V	Sub W
Average SoP($, £8)	4.02	3.91
Average SoP(P, £8)	3.69	2.71
Average SoP($, £4)	4.77	4.81
Average SoP(P, £4)	5.60	4.39
Average SoP($, R0.20)	3.11	3.02
Average SoP(P, R0.20)	3.86	3.61
Average SoP($, R0.10)	4.46	4.58
Average SoP(P, R0.10)	4.36	4.23
[SoP($, £8) − SoP(P, £8)] > [SoP($, £4) − SoP(P, £4)]	52	48
[SoP($, £8) − SoP(P, £8)] = [SoP($, £4) − SoP(P, £4)]	7	7
[SoP($, £8) − SoP(P, £8)] < [SoP($, £4) − SoP(P, £4)]	10	14
[SoP($, R0.20) − SoP(P, R0.20)] > [SoP($, R0.10) − SoP(P, R0.10)]	13	11
[SoP($, R0.20) − SoP(P, R0.20)] = [SoP($, R0.10) − SoP(P, R0.10)]	5	4
[SoP($, R0.20) − SoP(P, R0.20)] < [SoP($, R0.10) − SoP(P, R0.10)]	51	54

SoP measures increasing preference for PR lottery from 1 to 8; 1–4 for comparator, 5–8 for PR lottery

tween £4 and each bet.[20] When the comparator is £8, the average SoP for the $-bet is higher than the average SoP for the P-bet in both subsamples, from which we infer a stronger average preference for the $-bet. However, when both bets are paired with £4, the inferred preference is the opposite for subsample V, where the average SoP for the P-bet is now nearly one category higher than for the $-bet. For subsample W, where the P-bet is substantially less attractive than the P-bet presented to subsample V, the average SoP for the P-bet remains lower than for the $-bet, but the difference has fallen from 1.20 to 0.42 categories.

However, comparisons with R-bets give a very different picture (see rows five to eight in Table 6). In these cases, it is the $-bet which is more similar to the comparators than either of the P-bets and so we see the average SoP($, R) responses changing much more than the average SoP(P, R) responses. In rows 5–8 of Table 6, we see that when the comparator is R0.20, the average SoPs for the P-bets are higher than for the $-bet in both subsamples, from which we infer a stronger average preference for the P-bet. However, when both bets are paired with R0.10, the inferred preference is reversed for both subsamples, contrary to the constant variance assumption that underpins simple scalability.

These patterns in the subsample means are also clearly evident at the individual level. Some relevant data are presented in the bottom half of Table 6. For both pairs

---

[20] We have coded the SoP categories so that they range from 1 to 8, with 1 representing the strongest SoP for the comparator and 8 representing the strongest SoP for the bet in question. We could have reported the SoP for all eight comparators, but taking just £4 and £8 from the certainties, and R0.20 and R0.10 from the R-bets is quite sufficient to illustrate the points we are making.

🙋 Springer

of comparators, we count cases in which the SoP difference favours the \$-bet more for the better comparator, cases in which it favours it to the same extent, and cases in which it favours it less. For example, in the ninth row of Table 6, we report cases in which [SoP(\$, £8) − SoP(P, £8)] > [SoP(\$, £4) − SoP(P, £4)], that is, as we move from £8 to £4 the relative SoP for the \$-bet is reduced. Asymmetric patterns are the rule. For certainties, 52 (48) subjects in subsample V (W) report relatively weaker preference for the \$-bet as the certainty is decreased from £8 to £4, while just 10 (14) do the opposite. By contrast, for the R-bets, 13 (11) subjects report relatively weaker preference for the \$-bet as we move from R0.20 to R0.10, while 51 (54) do the opposite. For all comparisons, if changes in the two directions were equally probable, the probability of such asymmetric splits happening by chance would be vanishingly small. So, our individual-level SoP data provide further evidence against the constant variance assumption of models that satisfy simple scalability.[21]

6 The Marshak-Machina pairs and the independence axiom

The remaining 28 tasks of the experiment involved 14 pairs of lotteries (with each pair presented twice) that can be represented in the Marschak-Machina (M-M) triangle (e.g. Machina 1982) shown in Fig. 6 below. We will refer to these as M-M pairs.

The vertical edge of the triangle shows the probability of the highest payoff, **h**, and the horizontal edge shows the probability of the lowest payoff, **l**. The probability of the middle payoff, **m**, is given by $1 − \text{pr}(\mathbf{h}) − \text{pr}(\mathbf{l})$. So the points A, B and C on the vertical edge represent progressively worse probability mixes of **h** and **m**: for example, C = (**h**, 0.25; **m**, 0.75). D is the certainty of **m**, while E, F and G are progressively inferior mixes of **m** and **l**. The points H to N on the hypotenuse all represent different mixes of **h** and **l**, becoming unambiguously worse as we move south-east along the hypotenuse. For both subsamples, we used **h** = £25 and **l** = 0: the only difference was the value of **m**, which was £12 for Subsample V and £8 for Subsample W. The fourteen ways in which prospects were paired are shown by the 14 lines connecting pairs of labelled points.

Pairs such as those in Fig. 6 have played a major role in tests of the independence axiom of EUT, which implies that an agent's preferences can be represented in the M-M triangle by a map of indifference curves taking the form of parallel straight lines. For any pair of prospects on a given line, the one to the south-west can be regarded as safer (S) than the riskier (R) one to the north-east. If EUT is interpreted as a deterministic model, the implication is clear: for all {S, R} pairs connected by straight lines with the same gradient, an expected utility maximiser should always choose S, or else always choose R. Under EUT, switching from S to R or from R to S can only occur as a result of some noise or imprecision in people's behaviour. Any asymmetric pattern between S to R (SR) and R to S (RS) switches that cannot be explained by a purely stochastic component in people's judgments is liable to be counted as a violation of the independence axiom.

---

[21] These results strengthen our conclusion, reported in Butler et al. (2012) on the basis of the choice data of the same tasks, that the constant variance assumption is inappropriate in decision problems of this kind.

Fig. 6 The M-M pairs
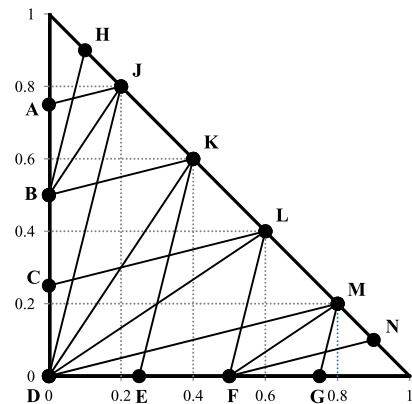
Many experiments have presented participants with pairs such as {D, J} and {F, L} in Fig. 6, where the second pair can be obtained by scaling the probabilities of the two highest payoffs down by a constant factor (0.5 in this example), thereby preserving the ratio between the two. Since at least Kahneman and Tversky (1979), the typical pattern in such cases has been for a larger percentage of participants to choose D from {D, J} than choose F from {F, L}, making the frequency of SR switches far larger than that of RS switches. Such a tendency has been dubbed the Common Ratio Effect (CRE) and has widely been interpreted as a systematic violation of EUT's independence axiom, motivating the development of various alternative theories which relax independence in one way or another. But how far the change in the proportions of safe choices really reflects non-EU preferences and how far it may be due to noise is an open question (see Loomes and Sugden 1995; and more recently Bardsley et al. 2009, Chap. 7).

As previewed in Sect. 2, this is an issue that SoP judgments may help us to address, especially if we also bring {B, H} into play. B and H are the same distance apart as F and L and are also connected by a line with a gradient of 4, which means that under EU assumptions there would be the same subjective value difference for {B, H} as for {F, L}, so that the direction and strength of preference should be the same for both of these pairs, while for {D, J} the direction of preference should be the same but the SoP should be substantially higher (note that {D, J} and {F, L} for subsample V correspond exactly to {S, R} and {S′, R′} in the example of Sect. 2). A parallel implication holds for {B, K}, {D, M} and {F, N}, where all three pairs are joined by lines with a gradient of 1/4 and where the distance between B and K is the same as that between F and N, with both being half of the distance between D and M. We can use the SoP responses to examine these implications.

The data produced in these pairs and in another eight pairs within the same triangle are reported in Table 7. The pairs are organised into groups joined by lines with the same slope. Each row reports the number of observations for each of the eight SoP categories obtained by assigning the mean SoP score recorded in the two repetitions of each task to the category corresponding to the verbal description that was shown

🖄 Springer

Table 7 SoP distributions in M-M pairs

Pair {S, R}	SoP categories based on average score								Sop		SoP range		S (%)
	S preferred				R preferred				Mean	Std. dev.	Mean	Std. dev.	
	vmb	mb	b	sb	sb	b	mb	vmb					
Subsample V													
Gradient = 4													
{B, H}	11	7	14	10	10	7	3	2	3.69	1.92	1.57	1.83	66 %
{D, J}	1	3	10	12	20	15	4	4	4.86	1.54	0.97	1.25	38 %
{E, K}	0	0	3	21	28	9	4	2	4.94	1.07	0.61	1.06	36 %
{F, L}	1	0	0	13	31	16	4	3	5.24	1.12	0.80	1.08	21 %
{G, M}	1	0	1	18	25	15	3	4	5.13	1.23	0.81	0.81	30 %
Gradient = 3/2													
{B, J}	14	12	17	13	6	2	1	2	3.07	1.69	1.35	1.43	84 %
{D, K}	5	8	19	15	16	2	3	0	3.69	1.45	1.04	1.33	69 %
Gradient = 2/3													
{D, L}	5	12	19	23	6	2	2	0	3.39	1.32	0.97	0.98	86 %
{F, M}	1	4	12	43	6	1	1	1	3.88	1.02	0.58	0.91	87 %
Gradient = 1/4													
{A, J}	26	20	10	8	2	1	2	0	2.29	1.47	0.99	1.53	93 %
{B, K}	20	24	16	3	3	0	1	1	2.32	1.39	1.09	1.34	93 %
{C, L}	11	25	22	7	2	1	1	0	2.58	1.19	1.30	1.25	94 %
{D, M}	11	10	27	15	3	1	0	2	3.03	1.43	0.75	1.02	91 %
{F, N}	3	4	23	32	5	0	0	2	3.61	1.17	0.67	0.87	90 %
Subsample W													
Gradient = 4													
{B, H}	4	4	9	9	17	15	8	3	4.78	1.78	1.30	1.45	38 %
{D, J}	1	1	5	10	20	18	13	1	5.29	1.37	0.90	1.14	25 %
{E, K}	0	0	1	7	30	23	5	1	5.40	0.89	0.67	1.01	12 %
{F, L}	0	0	0	2	32	20	9	4	5.72	0.95	0.83	0.79	3 %
{G, M}	0	0	0	5	29	23	9	2	5.62	0.91	0.59	0.75	7 %
Gradient = 3/2													
{B, J}	3	10	23	16	10	4	2	1	3.65	1.44	1.22	1.44	75 %
{D, K}	2	6	8	20	21	8	3	0	4.29	1.36	1.07	1.09	53 %
Gradient = 2/3													
{D, L}	1	6	10	38	10	3	1	0	3.91	1.04	0.58	0.79	80 %
{F, M}	1	0	5	42	17	4	0	0	4.25	0.79	0.64	0.82	70 %
Gradient = 1/4													
{A, J}	14	21	17	12	3	0	2	0	2.67	1.36	1.04	1.39	93 %
{B, K}	17	10	22	17	1	1	1	0	2.74	1.32	0.88	1.37	96 %
{C, L}	6	14	31	11	5	2	0	0	3.01	1.13	1.14	1.23	90 %
{D, M}	2	6	22	30	8	1	0	0	3.57	0.96	0.49	0.70	87 %
{F, N}	1	2	15	42	6	3	0	0	3.86	0.84	0.59	0.73	87 %

underneath the SoP bar.[22] The remaining five columns report, respectively: the mean and standard deviation of SoP for each pair measured on a scale from 1 to 8, where 1 represents strongest preference for S, and 8 strongest preference for R; the mean and standard deviation of SoP range, that is, the number of SoP categories between the two presentations of the task; and the percentage of the subsample whose SoP reveals a preference for the safer lottery in the pair.

These 14 pairs allow multiple within-sample comparisons to see how the distributions of SoP responses alter as one prospect is held constant while the prospects with which it is paired are varied. For example, comparing {B, H} with {B, J} and {B, K} should produce distributions which move in favour of B as the alternative on the hypotenuse gets steadily worse. Likewise, comparing {E, K}, {D, K} and {B, K}, we should expect the distributions to move against K and in favour of the safer options as these improve; and that is what we see. Other comparisons tell a similar story: ESM provides further details of the within-sample responsiveness of the SoP distributions in the expected direction.

There is also substantial evidence of between-sample differences consistent with the fact that the middle payoff was higher in subsample V than in subsample W. The final column in Table 7 shows that the percentage of participants who prefer the safer option was higher in V than in W for 12 out of 14 pairs, equal in one pair and slightly lower in one pair, with these last two involving cases where at least 93 % of participants prefer the safer option. Further to the left, the SoP Mean column of Table 7 shows that members of subsample V favour the safer option more strongly in *every one* of the 14 comparable pairs:[23] in this sense, the SoP data exhibit sharper between-sample sensitivity than pairwise choice proportions alone.

Turning specifically to the triples {B, H}, {D, J}, {F, L} and {B, K}, {D, M}, {F, N}, we see clear evidence at odds with EUT's independence axiom. For the first of these triples, the violations can be seen in binary choices alone, with the proportions of S choices reducing strongly and progressively as we move from {B, H} to {F, L}. The SoP responses provide further reinforcement: for example, even those B-choosers who stuck with the safer D option tended to give lower SoP responses, despite the distance between D and J being double the distance between B and H.

However, for the pairs joined by a gradient of 1/4, binary choices alone do not tell us much, since the overwhelming majority of participants choose S in all three pairs.[24] Yet it is clear from Table 7 that, for both subsamples, SoP reduces greatly as we move from {B, K} to {D, M} even though EU implies it should rise; and is *much* lower for {F, N} than for {B, K} even though it should, according to EU, be the same.

---

[22]For these tasks too we investigated whether there were any systematic trends in the patterns of answers between the first and second presentation of the questions. We used a test analogous to the one we used with the PR pairs. In subsample V, the average SoP did not change significantly between the first and second presentation, while in subsample W there was a tendency, in some of the pairs, for the riskier option to be more strongly preferred when the same task was faced for the second time.

[23]Recall that in our coding system, stating that the safer option is vmb is coded as 1, stating that the safer option is mb is coded as 2, and so on through to assigning 8 to cases where the riskier option is stated to be vmb. Thus higher Mean SoPs signify shifts in responses, either reducing the SoP for the safer option in any pair, or increasingly favouring the riskier option, or both.

[24]This imbalance in preference direction is not a bias against observing the CRE *per se*, as would have been the case had preferences in that pair been strongly in favour of the risky option.

There is something else that the SoP data allow us to look at: namely, the amount of noise surrounding the various choices. Since we have two SoP responses from each participant for every binary choice, we can derive a measure of the noise associated with any pair by computing the distributions of individuals' differences between those two SoP responses. Such measures are reported in the third to last column of Table 7. If the judgmental noise is much the same for all pairs, we should not see any significant trends in the distributions of the differences from one pair to another. But if the variance of the noise term is some function of the magnitudes of the subjective values of the prospects and/or the size of the SoP, we might expect to see this reflected by changes in those distributions. What we find is that there is more noise for pairs such as {B, H} and {B, K}, which are associated with both greater subjective values and larger SoP responses. These few comparisons alone do not allow us to disentangle the contributions of different factors to the noise, but the extended use of SoP responses might help us to do so. In line with our findings in the rest of the paper, our SoP data suggest that the assumption that the error term has constant variance may be highly questionable.

7 General discussion

Our study was motivated by a desire to explore decisions in a richer way than is possible with choice-only data. As we have seen, presenting the same preset alternatives to samples of heterogeneous individuals and relying on binary choices alone can give an inadequate and sometimes misleading picture of the phenomena of interest. In an attempt to address these issues, we have tried to produce more fine-grained data by eliciting self-reported SoP judgments. In this section, we reflect on the main limitations of our instruments, and outline ways in which it may be improved.

A first possible reason for concern is the inherent discretisation of our data, highlighted by the spikes in Fig. 4. If the objective is to get richer information about preferences, this aspect may be seen as a disappointment, as the resulting data may not be as rich as one would like them to be for certain kinds of applications. For our instrument, that discretisation seems to be a direct result of changing the wording at various points along the slider. A possible solution would be to avoid the wording altogether and just anchor the ends of the SoP scale. We have experimented with something similar in other studies in which we have measured the SoP for the chosen option on an 11-point Likert scale anchored so that 11 points represent the difference between what participants regarded as the best and the worst lotteries in the set from which pairs were constructed. The general finding is that many participants' ability to discriminate seems confined to a subset of the scale provided, in line with results from cognitive psychology (e.g. Miller 1956).[25]

So would it be possible to use a different metric? In other research, we have experimented with a candidate that many would regard as appealing: money. If SoP could be measured in units of money, several problems would be solved at the same time: the resulting measures would be meaningful to participants (and researchers); they

---

[25] These studies are still unpublished. More details are available from the authors upon request.

would provide continuous data on a ratio scale; and it would be possible to directly link them to monetary incentives. Unfortunately, our extended efforts in this direction have highlighted some serious difficulties with such instruments, which result in systematic discrepancies between monetary SoP measures and choice behaviour.[26]

A second possible reason for caution relates to the difference between SoP and *confidence*. While the two are theoretically distinct concepts, they may not be very easy to disentangle. Conceptually, SoP refers to what we have called the central tendency difference between the subjective values of the options, whereas confidence reflects the strength of the participant's belief that he has chosen the option that really is best for him, with this 'strength of belief' judgment depending at least in part on the noisiness of his preferences.

In our dominance pairs, it was quite easy for respondents to identify the option that was unambiguously better, so that they could be completely confident about their choice even when the incremental difference was small. The high frequency of maximal SoP responses by some respondents in such cases suggests that those individuals may have confounded confidence with SoP. We have found in subsequent studies that an introductory example, illustrating the distinction between the two concepts, seems to help participants isolate SoP more effectively in dominance pairs.

There are other cases, however, in which SoP and confidence may be more tightly intertwined. Consider a case in which two alternatives are finely balanced, so that the central tendency difference between their subjective values is close to zero, and therefore the probability that one is chosen over the other is close to 0.5. In this case, both SoP and confidence are low. More generally, in cases where neither option dominates, SoP and confidence may often be quite highly correlated.[27] Making progress in disentangling these key constructs may require a theoretical framework that moves beyond standard economic analysis to include aspects inherent in the *process* through which decisions are made. Psychologists have already produced process models that make explicit predictions about confidence (e.g. Pleskac and Busemeyer 2010), which may illuminate the distinction and perhaps suggest ways of eliciting the two constructs separately. But in undertaking such an exercise, a further complication should be borne in mind: once it is accepted that choice behaviour is characterised by a considerable amount of noise, it must be expected that SoP and confidence judgments will also be noisy in their own ways, making it even more difficult to separate them.

Before concluding, we turn to another important issue in taking this research programme forward. Can SoP judgments be used to improve fitting-and-prediction exercises? As we have noted repeatedly, a particularly crucial challenge in this respect is to get the stochastic specification right, as the wrong specification may misidentify people's core preferences and lead to incorrect predictions. Since SoP data are measured on a richer scale than choice data, they may have the potential to give more insights with fewer observations.

---

[26]Our attempts to elicit SoP on a money scale are documented in Butler et al. (2013), where we report what appear to be strong and systematic biases.

[27]In other experiments, we have elicited both confidence (and sometimes decision *difficulty*) and SoP, and found them to be highly correlated.

⚫ Springer

8 Concluding remarks

Our study had two main objectives: testing whether our SoP instrument produces sensible data despite the lack of monetary incentives; and illustrating the additional insights that can be obtained from the extra degree of granularity. How far have these objectives been achieved? We have shown that the SoP responses display considerable sensitivity to variations in the parameters of the lotteries, not only at the within-subject level—the one that is essential for the instrument to be informative in most of the applications we have considered—but also at the between-sample level. This latter result provides grounds for thinking that what the instrument is tapping into is something which is meaningful to most subjects and which appears to travel quite well across different sets of binary choices, allowing broad comparability. The fact that almost all individuals use the full range of the bar (rather than simply nudging it a minimal amount to indicate choice, or pushing it straight to one end or the other) suggests that intrinsic motivation, activated by a simple request, is sufficient to yield a great deal of useful additional information about preferences.

We have also provided examples of how one can use the information provided by SoP judgments. By looking at the variability of these judgments when the same task is repeated a number of times, we can learn something about the stochastic component of preferences. SoP judgments can be used to investigate whether subjective value differences behave as implied by simple scalability, or whether properties such as EUT's independence axiom hold at the individual level. On balance, we think that there is a case for SoP judgments to be a useful addition to an experimenter's toolbox.

As we have highlighted in the previous section, there are various ways in which our SoP instrument might be improved. And there may be other measures that could further supplement our datasets: for example, individuals' judgments of the difficulty of deciding, their confidence that they have made the right decision, and the response times involved, which are regarded as key elements in the understanding of decision-making processes (see Pleskac and Busemeyer 2010). Such measures may or may not turn out to be robust and useful—that remains to be seen—but if they can be shown to pass various consistency checks, they may, in conjunction with choices and SoP judgments, help to provide a multi-faceted description of behaviour that can improve our understanding and our models. Moreover, if it turns out that such measures are robust and useful in the context of individual decision making under risk, it may be possible to extend their use to other areas of economics such as intertemporal decisions or the study of beliefs and actions in strategic situations.

We acknowledge that there is a strong tradition in experimental economics of relying as much as possible on data that are collected under appropriate incentive conditions, and we subscribe to this rule as a general principle. However, that does not mean that other kinds of data are necessarily valueless. When other potentially informative data cannot easily be linked to material incentives, we can still check whether they conform with certain basic criteria of coherence and consistency; and if they do, they may help us better understand decision processes and the resulting patterns of behaviour.

🌲 Springer

References

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2009). *Experimental economics: rethinking the rules*. Princeton: Princeton University Press.

Becker, G., DeGroot, M., & Marschak, J. (1963). Stochastic models of choice behaviour. *Behavioral Science*, *8*, 41–55.

Blavatskyy, P. R. (2011). A model of probabilistic choice satisfying first-order stochastic dominance. *Management Science*, *57*, 542–548.

Buschena, D., & Zilberman, D. (2000). Generalized expected utility, heteroscedastic error, and path dependence in risky choice. *Journal of Risk and Uncertainty*, *20*, 67–88.

Busemeyer, J. R. (1985). Decision making under uncertainty: simple scalability, fixed sample, and sequential sampling models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *11*, 538–564.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Butler, D., & Loomes, G. (2007). Imprecision as an account of the preference reversal phenomenon. *The American Economic Review*, *97*, 277–297.

Butler, D., Isoni, A., & Loomes, G. (2012). Testing the 'standard' model of stochastic choice under risk. *Journal of Risk and Uncertainty*, *45*(3), 191–213.

Butler, D., Isoni, A., Navarro-Martinez, D., & Loomes, G. (2013). *On the measurement of strength of preference in units of money*. Unpublished Manuscript.

Fechner, G. T. (1860). *Elemente de psychophysik*. Amsterdam: Bonset. (Reprinted in 1966 by Holt, Rinehart and Winston, New York).

Hey, J., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*, 1291–1326.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*, 46–55.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648.

Luce, D. R., & Suppes, P. (1965). Preference, utility and subjective probability. In D. R. Luce, R. R. Bush, & E. H. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 249–410). New York: Wiley.

Machina, M. (1982). 'Expected utility' theory without the independence axiom. *Econometrica*, *50*, 277–323.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*, 371–404.

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: a theory of confidence, choice, and response time. *Psychological Review*, *117*(3), 864–901.

Seidl, C. (2002). Preference reversal. *Journal of Economic Surveys*, *16*, 621–655.

Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*, 332–382.

Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology*, *6*, 1–12.

Von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge: Cambridge University Press.

Wilcox, N. T. (2011). 'Stochastically more risk averse': a contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89–104.