

**Original citation:**

Loomes, Graham and Pogrebna, Ganna. (2014) Testing for independence while allowing for probabilistic choice. *Journal of Risk and Uncertainty*, Volume 49 (Number 3). pp. 189-211. ISSN 0895-5646

**Permanent WRAP url:**

<http://wrap.warwick.ac.uk/63290>

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 3.0 (CC BY 3.0) license and may be reused according to the conditions of the license. For more details see: <http://creativecommons.org/licenses/by/3.0/>

**A note on versions:**

The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: [publications@warwick.ac.uk](mailto:publications@warwick.ac.uk)



<http://wrap.warwick.ac.uk>

# Testing for independence while allowing for probabilistic choice

Graham Loomes · Ganna Pogrebna

Published online: 23 December 2014

© The Author(s) 2014. This article is published with open access at Springerlink.com

**Abstract** We propose a broad framework for individual choice under risk which can accommodate many stochastic formulations of various deterministic theories. Using this framework to guide an experimental design, we show that most individuals' departures from the independence axiom cannot be explained by adding a 'random noise' term to a deterministic 'core' theory which incorporates this axiom. We also find behaviour that cannot be explained in terms of the standard assumptions of Cumulative Prospect Theory, often invoked to account for violations of independence. Our results suggest that 'similarity' effects may explain the data better.

**Keywords** Independence axiom · Stochastic choice · Expected utility theory · Prospect theory

**JEL Classifications** C81 · C91 · D81

## 1 Introduction

Expected Utility Theory (EUT) and many of the most influential alternative decision models developed during the past 40 years are normally expressed in deterministic form.<sup>1</sup> However, when we wish to fit them to data and/or seek to use experimental

---

<sup>1</sup> There are now many such alternatives: Prospect Theory (Kahneman and Tversky 1979); various rank-dependent expected utility models (e.g., Quiggin 1982) and the reformulation of prospect theory in rank-dependent terms (Tversky and Kahneman 1992); Machina's (1982) Generalized Expected Utility Theory; Regret Theory (Bell 1982; Loomes and Sugden 1982); Disappointment Theory (Bell 1985; Loomes and Sugden 1986); similarity models (Rubinstein 1988; Leland 1994); and more besides—see Starmer (2000) and Sugden (2004) for reviews—although more variants have appeared since. We shall refer to models expressed in deterministic form as 'core' theories.

---

G. Loomes (✉)

Warwick Business School, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK  
e-mail: g.loomes@warwick.ac.uk

G. Pogrebna

Warwick Manufacturing Group, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK  
e-mail: g.pogrebna@warwick.ac.uk

evidence to test and discriminate between them, we have to make appropriate allowance for the stochastic element in those data.

This presents us with a challenge, because there are a number of ways in which the stochastic component might be specified; and the parameters we estimate and the conclusions we draw from our tests may vary considerably depending upon the particular specifications used (for a discussion of these issues, see Stott 2006; Blavatsky and Pogrebna 2010; and Chapter 6 in Bardsley et al. 2010).

Rather than try to work through a multiplicity of somewhat differentiated variants, the strategy of this paper is to use a formulation that is sufficiently general that it is compatible with a broad spectrum of combinations of deterministic core theories and stochastic specifications. Within that framework, we report an experiment that allows us to discriminate between at least some different combinations. As we shall explain, it is not practicable to run the ‘ideal’ experiment, so the results and the strength of the conclusions we can draw are necessarily limited by the practical constraints. Nevertheless, we shall offer a design that others may adapt and extend in ways which will build up a body of evidence that will allow further discrimination between, and refinement of, descriptive decision models.

Meanwhile, the conclusions we draw from the experiment reported here can be summarised as follows:

1. The great majority of participants *do* display some degree of seemingly stochastic variability when making choices between simple lotteries: this underlines the need to develop probabilistic models and experimental designs capable of investigating them.
2. For many of these individuals, the patterns of response appear to diverge systematically from any model that requires (the equivalent of) the independence axiom at its core.
3. It is not possible on the basis of these data alone to discriminate fully between all of the various alternative core models that dispense with independence in one way or another. However, there are indications that although Cumulative Prospect Theory (Tversky and Kahneman 1992) has been widely adopted as the best alternative to EUT, it is not well-supported by the data in this experiment. These data point instead to the possibility that models which allow ‘similarity’ effects may organise the data better.

The paper is arranged as follows. In the next section we set out the general framework and discuss the issues we aim to examine. In Section 3 we describe the experimental design and the more specific questions it is intended to address. Section 4 reports the results and our interpretation of them. The final section offers some concluding remarks.

## 2 Modelling framework

Ever since the earliest attempts to use experiments to elicit people’s preferences over risky alternatives, it has been clear that many people’s choices behave probabilistically rather than deterministically. That is to say, 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,<sup>2</sup> he/she may answer differently in at least some of those repetitions.

A striking early example was provided by Mosteller and Nogee (1951). Their Fig. 2, reproduced below as our Fig. 1, plotted one individual's responses to a task involving 14 repetitions of 7 different variants of a particular decision task where the individual could either accept or refuse a bet involving a  $\frac{1}{3}$  chance of losing 5c and a  $\frac{2}{3}$  chance of winning an amount shown on the horizontal axis—either 7c, 9c, 10c, 11c, 12c or 16c.

When the winning payoff was 7c, this individual never accepted the bet, whereas when the winning sum was 16c, he always took the gamble. In between those two amounts, the frequency with which he took the gamble was greater than 0% and less than 100% and increased monotonically with the size of the prize.

So, without being specific about the possible source(s) of 'noise' in people's responses,<sup>3</sup> we take as a starting point a model where an individual's underlying preferences can be represented in the form of probability distributions and where it is as if each choice is an observation drawn from such a distribution.

To adapt the idea in Mosteller and Nogee's (1951) figure, and to provide a foretaste of the experimental design in Section 3, consider a particular lottery B which offers a 0.8 chance of £40 and a 0.2 chance of zero, written briefly<sup>4</sup> as  $B=(£40, 0.8)$ ; and consider choices between B and various sure alternatives  $A=(x_j, 1)$  where  $x_j$  increases in £2 steps from  $x_0=£14$  to  $x_{13}=£40$ . In Fig. 2, the deterministic preferences of an individual who is indifferent between B and a sure £26 are shown by the dashed step function: for all  $x_j < £26$ , she will always choose B and for all  $x_j > £26$  she will never choose B, while for  $x_j = £26$ , all probability mixes of A and B are equally good.

The continuous curve shows the underlying distribution for an individual with probabilistic preferences. In this particular example, when the sure thing offers £20 or less, the individual always chooses B, whereas when the sure thing offers £34 or more, he never chooses B. But when A offers amounts between £20 and £34, there is some chance that either A or B might be chosen. For example, when presented with a choice between B and £22 for sure, there is a 0.95 chance he will choose B; so if he made that choice 100 times on separate and independent occasions, we would expect on average to observe him choosing the sure £22 on five of those occasions. Increasing the sure amount to £30 greatly increases the likelihood that the sure sum will be chosen—to 93%—but there is still a 0.07 chance of choosing B.

<sup>2</sup> In some experiments, the repetitions occur within a single experimental session. Other experiments have asked participants to attend two or more sessions on different days within the same week, or perhaps over several weeks. An advantage of spreading the repetitions over several days or weeks is that participants may forget earlier answers and therefore treat each 'round' more freshly and independently. But there is generally some attrition and possibly changes in people's circumstances between sessions such that variation in their responses is not random. There appears as yet to be no good evidence about the balance of advantage and disadvantage in this respect.

<sup>3</sup> In saying this, we are not suggesting that the source of variability in individuals' responses is unimportant: on the contrary, we think it will be an important question to address as part of the longer-term development of descriptive models of human decision making. However, for our present purposes it is not necessary to identify any particular explanation.

<sup>4</sup> In this paper, all lotteries involve one positive payoff and zero otherwise: for brevity we omit the zero.

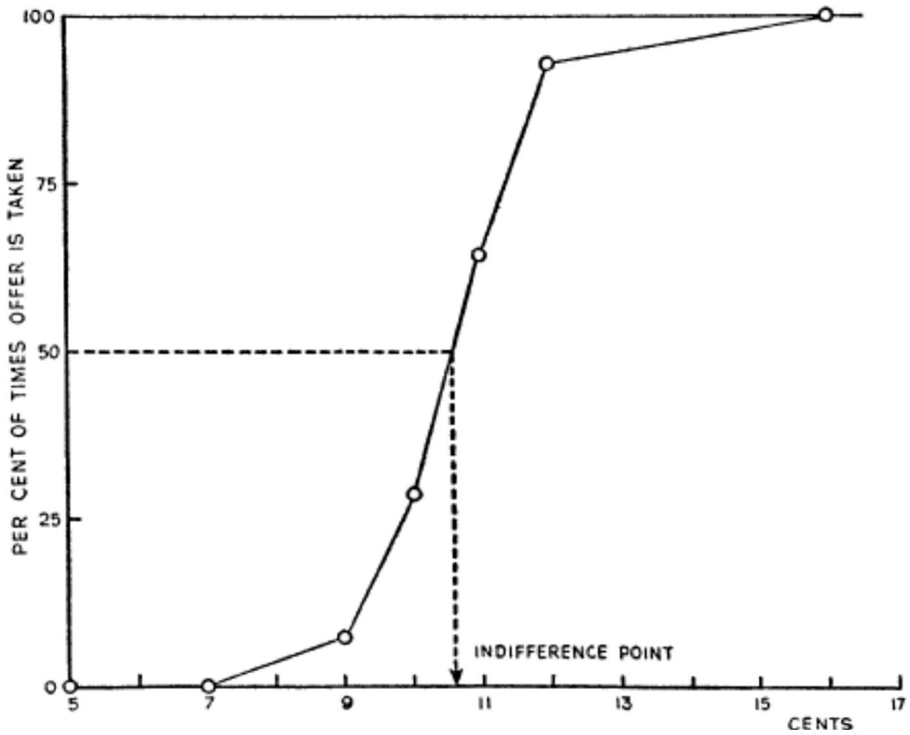


Fig. 1 Diagram from Mosteller and Noguee (1951)

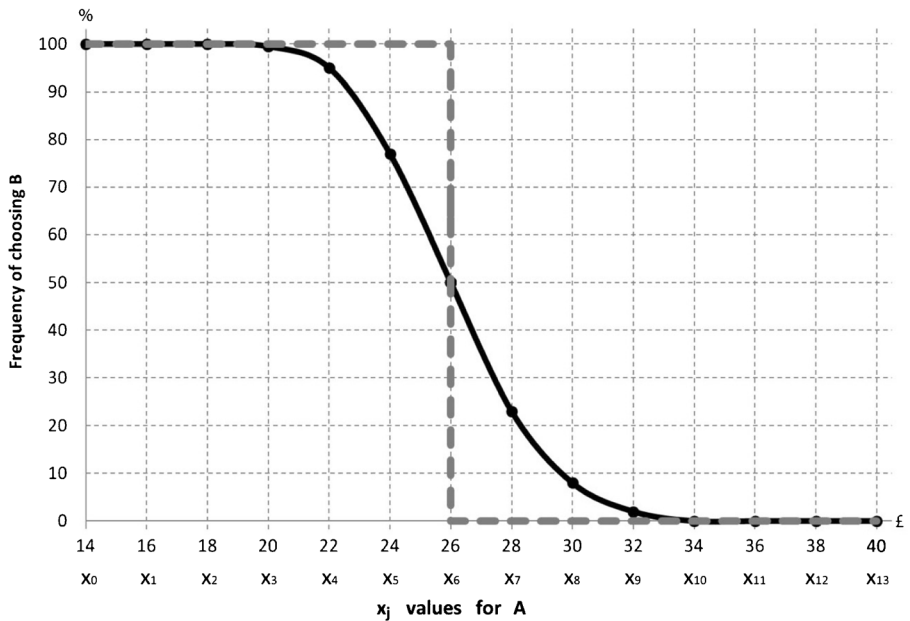


Fig. 2 Deterministic and probabilistic preferences

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). In the example shown in Fig. 2, the SI point is where A offers £26 for sure.

There are various different ways in which we might ‘explain’ Fig. 2. At this stage, we shall not go into detail about these different accounts, but simply note that all of them are broadly consistent with the general pattern depicted in our Fig. 2.<sup>5</sup> Nor should anyone read anything very specific into the Fig. 2 illustration: the curve there happens *not* to be symmetrical around the SI point, but at this stage we are not making any strong assumptions either about symmetry or about the nature of any asymmetry. Figure 2 happens to take a particular lottery as fixed and compares it with increasing sure sums ranging along the horizontal axis; but we could also plot a function where the A options are non-degenerate lotteries offering steadily improving prospects—for example, we might have supposed  $A = (£100, p)$ , with  $p$  increasing from (say) 0.1 by increments of 0.05 to 0.75. In this latter case, the curve might well take a different shape from the one shown in Fig. 2, but we might still not be surprised to find some lower values of  $p$  such that B is always chosen, followed by some range where B is progressively less likely to be chosen as  $p$  increases until we reach some higher values of  $p$  such that B is never chosen.

However, taking the A options in Fig. 2 to be sure sums of money is convenient for our current exposition because it allows us to consider how we might represent another series of comparisons between lotteries C and D that are generated from A and B. Let us take B and scale down the probability of the £40 payoff by a quarter in order to generate  $D = (£40, 0.2)$ ; and let us take any A and scale down the probability of the sure payoff by a quarter and substitute extra chances of zero in order to generate  $C = (x_j, 0.25)$ . According to the independence axiom of deterministic EUT, whatever preference an individual has for  $\{A, B\}$ , he should have the corresponding preference for  $\{C, D\}$ —replacing B by D and replacing the A options by the corresponding C lotteries should leave the step function exactly as it is in Fig. 2.

But is the same true for a probabilistic formulation with EUT as its core? To begin to answer this question, we need to be more specific about what is meant by ‘a probabilistic formulation with EUT as its core’.

At this stage, it may be sufficient to think in terms of a ‘strong’ formulation and a ‘weak’ formulation. A strong formulation would entail the *whole distribution* shown in Fig. 2 being applied exactly as it is to the choices between D and different Cs. Various stochastic specifications of EUT have this implication (see Luce and Suppes 1965); but a weaker formulation requires only that the distribution for  $\{C, D\}$  has the same SI point as the distribution for  $\{A, B\}$ . Weaker specifications allow the rest of the function to be configured differently, as in Fig. 3 where the solid line depicts the probability function for  $\{C, D\}$  while the dashed line superimposes the  $\{A, B\}$  distribution from Fig. 2. Such specifications require only that  $\Pr(A \succ B) \geq 0.5 \leftrightarrow \Pr(C \succ D) \geq 0.5$ .

Consider a particular case where  $x_j = £30$  so that  $A = (£30, 1)$  and  $C = (£30, 0.25)$ . In this case, the two pairs  $\{A, B\}$  and  $\{C, D\}$  in our example constitute a format popularised by Kahneman and Tversky (1979) to demonstrate what they called the

<sup>5</sup> A thorough review of different variants of probabilistic models extant at the time was provided by Luce and Suppes (1965). A more recent discussion can be found in Rieskamp et al. (2006). Most of the models covered in these reviews are compatible with the representation in Fig. 2.

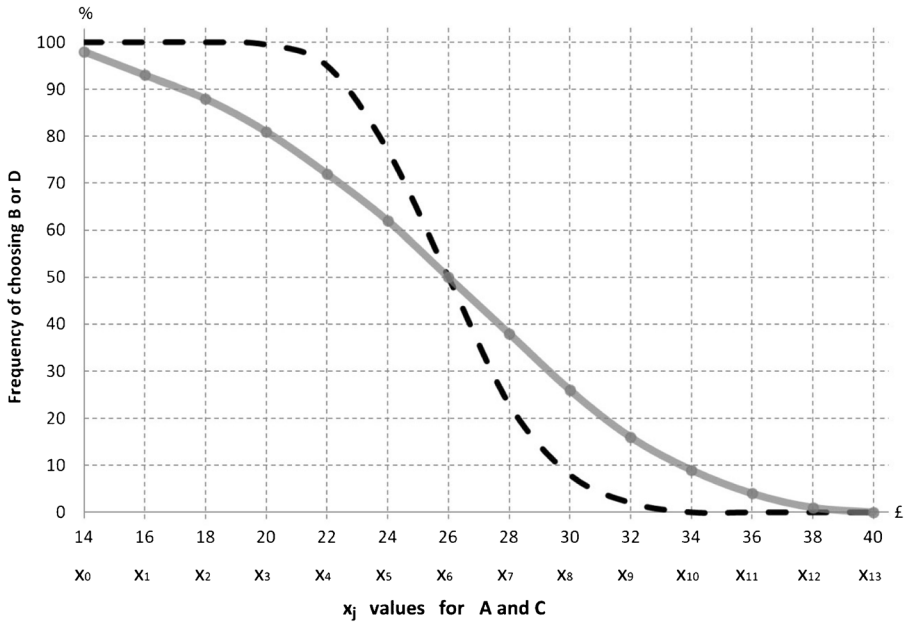


Fig. 3 Possible distributions for ‘Scaled-Up’ and ‘Scaled-Down’ Lotteries

common ratio effect (CRE). In terms of deterministic EUT, there are two ways of ‘violating’ the independence axiom in this example: namely, choosing the safer option A in the scaled-up {A, B} pair while also choosing the riskier option D in the scaled-down {C, D} pair; or else choosing the riskier B from the first pair while also choosing the safer C from the second pair. The typical pattern found by Kahneman and Tversky and replicated in many other experiments undertaken during the past 35 years is that the A & D violation is exhibited by substantially more people than the B & C violation. This asymmetry has been an influential part of the evidence against EUT as an adequate descriptive model.

Figure 3 shows how such an asymmetry *might*—at least, up to a point—be compatible with a probabilistic formulation of EUT. If (as in our example above) we set the safer payoff at £30, we can see that  $\Pr(A > B)=0.93$  while  $\Pr(C > D)=0.74$ , so that if each choice is made separately and independently,  $\Pr(A \& D)=0.2418$  while  $\Pr(B \& C)=0.0518$ . Translated to a sample of 100 individuals with probabilistic preferences like those in Figs. 2 and 3, we would have some 24 ‘violations’ of the A & D kind compared with just five of the B & C kind—an asymmetry that might appear at first sight to be systematic and unlikely to have arisen by chance. The ‘unlikely-to-have-arisen-by-chance’ judgment is often made on the basis of a null hypothesis that both patterns ‘ought’ to be equally likely. But that is not the correct null in this case. It *would* be the correct null for the ‘strong’ formulation that entails the *whole distribution* being the same for both choice pairs; but as shown in relation to Fig. 3, such an asymmetry could be entirely compatible with the ‘weak’ probabilistic formulation of EUT.

However, notice two other implications of the weaker formulation depicted in Fig. 3. First, there is a limit to the extent of the CRE pattern. If the SI point is the same for {A,

B} as for {C, D}, it cannot be that  $\Pr(A \succ B) > 0.5$  while  $\Pr(C \succ D) < 0.5$ : that is, for a sample of individuals like the one in the example, we should not expect to see a reversal of the *modal* choice of the kind that has been reported in at least some studies, including Kahneman and Tversky (1979).

Second, if Fig. 3 were an adequate representation of a typical individual's preferences, it should be possible to set the parameters in such a way as to produce a pattern which shows the opposite asymmetry. Consider the case where  $x_j$  is set at £22 and label the alternatives to B and D as  $E=(£22, 1)$  and  $F=(£22, 0.25)$ . Reading off from Fig. 3, this would entail  $\Pr(E \succ B)=0.05$  while  $\Pr(F \succ D)=0.28$ , so that  $\Pr(E \& D)=0.036$  while  $\Pr(B \& F)=0.266$ —an even greater asymmetry, but now the other way round. Such 'reverse-CRE' patterns have seldom been reported<sup>6</sup>—although, for exceptions, see Blavatsky (2010) and Loomes (2010).

On the other hand, how might things look if the probabilistic model has a core which dispenses with the standard independence axiom? In particular, suppose we take one of the deterministic models where those who are indifferent between A and B have a strict preference for D over C. One possibility is shown in Fig. 4, where the dashed line shows the original distribution from Fig. 2 with the SI point for {A, B} at  $x_j=£26$ , while the solid line depicts a distribution for the {C, D} pair where the SI point occurs when  $x_j$  is just under £31. In this particular example, when the solid line is between 100% and 0% it is everywhere above the dotted line. This would allow the modal choice to switch for all values of  $x_j$  between £26 and just under £31. But in this example there would never be any reverse-CRE patterns: even at lower levels of  $x_j$  such as £22 and £24 the asymmetry would be in the usual direction.

Of course, other combinations of non-EU core and probabilistic preferences are possible. For example, if the SI point for the {C, D} pair were when  $x_j$  is £28, and if the curve were generally flatter, as in Fig. 5, we could produce the standard CRE with a switch of mode when  $x_j$  lies between £26 and £28 and asymmetries of the usual kind when  $x_j$  is greater than £28. However, there would be no effect either way at  $x_j=£24$ ; and reverse-CRE patterns would be liable to occur at lower values of  $x_j$ .

All in all, then, 'snapshots' involving a particular value of  $x_j$ —with that particular value possibly being selected by the experimenter in order to produce a strong effect of a particular kind—may not give us sufficient information to distinguish between alternative models, especially if each choice is presented only once to each individual. It would be useful to try to get some richer picture of the kinds of distributions depicted in the Figures discussed and to see how/whether their shapes and/or locations change as we alter the parameters of the lotteries. In the next section, we describe how we attempted to do that.

<sup>6</sup> To some extent this lack of evidence might be due to some selection bias, of two kinds. First, those who wish to reproduce the standard effect in order to study it may choose the parameters most likely to generate it. Second, it might be that those who fail to produce the 'expected' effect are deterred from submitting their work for publication or are more likely to have it rejected. For either of these reasons—or perhaps simply because the reverse-CRE really doesn't occur very often—it is widely believed that the 'usual' CRE is a very robust regularity.



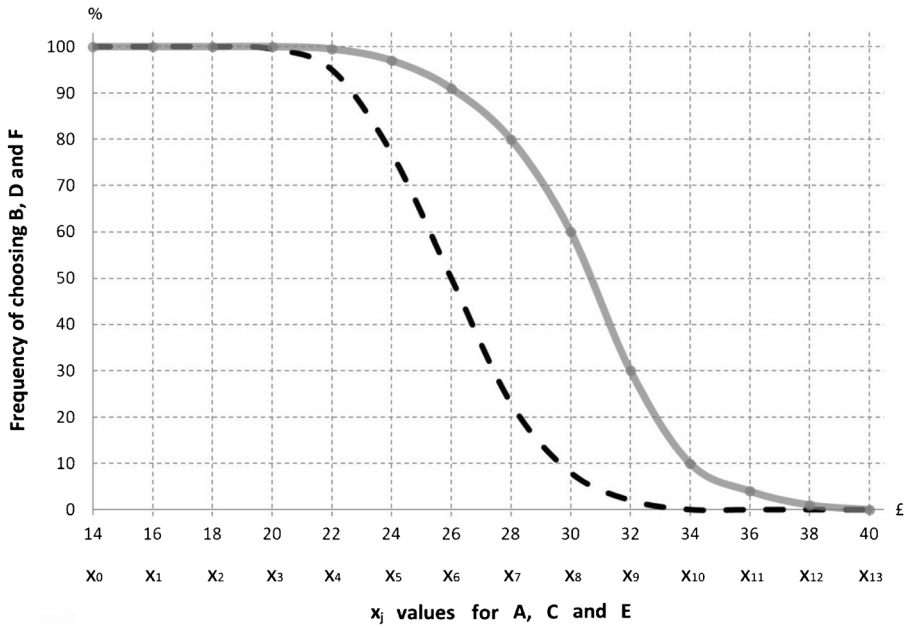


Fig. 4 Probabilistic preferences for a non-EU core

### 3 Experimental design

In principle, for any given B we should like to identify the range of  $x_j$ s that would cover the space from the highest  $x_j$  where B is still chosen with probability (very close to) one

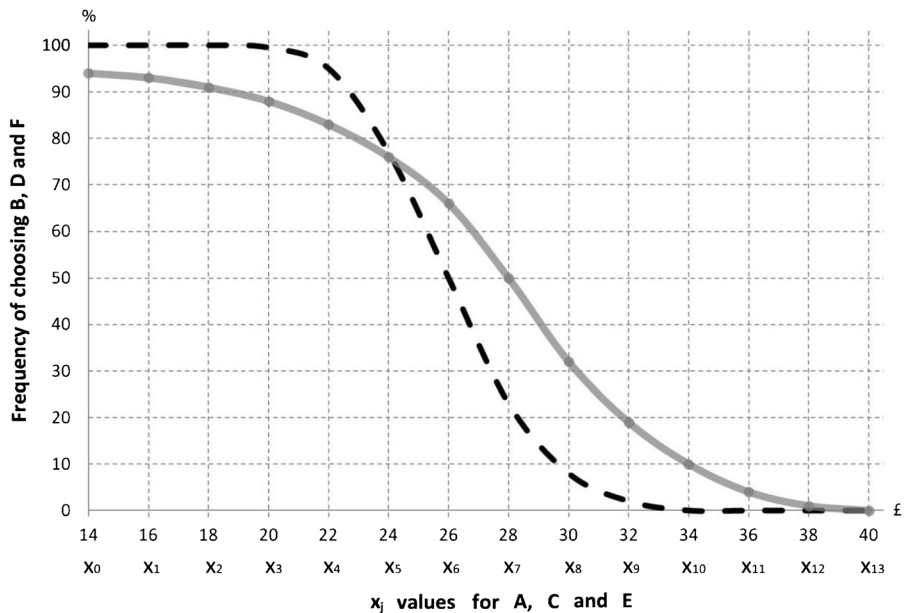


Fig. 5 A non-EU core with a flatter distribution

across to the lowest  $x_j$  where B is (almost) never chosen.<sup>7</sup> Within that range, we should ideally like to have enough {A, B} pairs each repeated enough times (with each choice independent of all earlier presentations of the same pair) to provide a good estimate of the curve in question.

Bearing in mind that any sample is likely to involve some degree of heterogeneity between individuals, one might like in this case to have at least 20 different  $x_j$  values no more than £1 apart with perhaps 10 repetitions of each pair, giving a total of 200 questions just for the {A, B} comparison. If we had the same number of questions for the {C, D} comparison, and if we interspersed a number of other rather different pairs to try to provide more variety and greater separation and independence between repetitions, this could easily raise the total number of questions to between 600 and 800. However, a design involving so many choices might be far too onerous and tedious to produce high quality data. Drawing on our own experience and advice from colleagues, we worked on the basis of a maximum of 200 questions, with just under half of these involving binary choices of the {A, B} and {C, D} form.<sup>8</sup>

Informed by a small pilot study, we opted for two ‘treatments’ to study the CRE, which we shall label CRESUMS (in which we varied the sums of money offered by the safer options) and CREPROBS (in which the probabilities in the riskier options were varied).

### 3.1 The CRESUMS questions

These questions aimed to obtain data about the kinds of choices discussed in relation to Figs. 2, 3, 4 and 5. That is, lottery B offered a 0.8 chance of £40 while the A options offered 11 different sure amounts from £16 to £36 inclusive, with £2 intervals between them. Likewise, lottery D offered a 0.2 chance of £40 while the C options offered the 11 different sums from £16 to £36 with probability 0.25. The 11 {A, B} and 11 {C, D} pairs were shown in random order, interspersed with other choices involving quite different parameters, and presented in four ‘waves’ within each experimental session so that every choice was made on four separated occasions. Pairs were displayed as in Fig. 6, where the upper panel shows a choice between B=(£40, 0.8) and A=(£30, 1), while the lower panel shows the corresponding scaled down pair with D=(£40, 0.2) and C=(£30, 0.25). We also reproduce the text that accompanied each display.

We opted for this way of displaying alternatives in order to try to strike a compromise between the ‘decision by description’ and ‘decision by experience’ approaches. A growing literature<sup>9</sup> suggests that when people form an estimate of probabilities on the basis of some sampling experience, they may behave differently

<sup>7</sup> Even in cases where one option transparently dominates the other, it has been observed that in a small proportion of cases—typically 1–2%—the dominated alternative is chosen. This is often regarded as a ‘pure’ mistake—a lapse of attention that results in the individual recording what he or she would normally agree is the opposite of what was intended. Our Figures omit such ‘trembles’—but we know that a small minority of responses will inevitably be of this kind.

<sup>8</sup> Details of the other questions are available from the authors. Also, those who wish to ‘participate’ in the experiment can access it via <http://gvp4c7.experimentaltoolbox.com/210> and <http://gvp4c7.experimentaltoolbox.com/211> for the two treatments of the experiment.

<sup>9</sup> A good selection is listed at <http://dfexperience.unibas.ch/literature.html>

**(a) B=(£40,0.8) versus A (£30,1)**

In this decision problem, we ask you to make a choice between an uncertain option and an amount of money for certain. The uncertain option involves some mix of red and black balls with each red ball worth £40 and each black ball worth 0, while the certain option is presented as 20 red balls each worth £30. If this decision problem is chosen for payment, we will look up which option (left or right) you chose. If you have chosen the amount of money for certain, you will receive that amount. If you have chosen the uncertain option, you will draw one ball from a bag with the appropriate mix of red and black balls and you will get paid the value of whichever colour is drawn.



**(b) D=(£40,0.2) versus C=(£30,0.25)**

In this decision problem, we ask you to make a choice between two uncertain options. Below you see a mix of red and black balls on each side. The difference is that on one side each red ball is worth £40 and each black ball is worth 0, while on the other side each red ball is worth £30 and each black ball is worth 0. If this decision problem is chosen for payment, we will look up which option (left or right) you chose. Then (dependent on your choice) you will draw one ball from a bag with the appropriate mix of red and black balls and you will get paid the value of whichever colour is drawn.

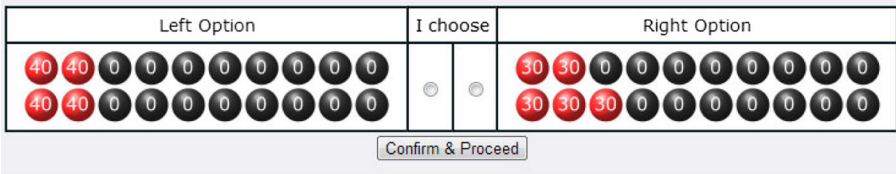


Fig. 6 Screenshots of questions used in the experiment

from when the probabilities are merely stated 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 showing 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 actually 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.

Some might think that such a format favours EUT (or any other model that supposes preferences to be linear in the probabilities), since the display may be regarded as emphasising such proportionality. However, the format does no more or less than show precisely what each option entails. If EUT is indeed the best model of people’s core preferences, there is every chance for it to perform well, whereas if some other model provides a better core, such a model should have the capacity to organise the data generated in this experimental environment.

In addition to the 44 binary choice questions for {A, B} and the corresponding 44 for {C, D}, we also presented participants with other binary choices involving very different parameters, and also different formats involving multiple price lists. These questions served to provide variety and separate the repetitions with which we are concerned in this paper.

### 3.2 The CREPROBS questions

The CRESUMS questions involve what might be thought of as ‘classic’ CRE parameters: a 0.8 chance of the high payoff versus an intermediate certainty in one pair, with both of those probabilities scaled down by a quarter for the other pair. We used those parameters in order to establish (a) whether the classic CRE does in fact occur under conditions of repeated binary choice between alternatives presented as in Fig. 6, and (b) whether the patterns of response are compatible with an EU core as in Fig. 3, or with some non-EU core as in Figs. 4 and 5.

One of the consequences of the seeming robustness of the CRE pattern of behaviour has been to stimulate theorists to produce a variety of models which account for those patterns in different ways—recall the examples given in footnote 1. We could not hope, within a single experiment, to investigate thoroughly all of the different possible accounts of the CRE; but we used the CREPROBS questions to look at two types of model that explain the pattern in ways that are sufficiently different that we can try to discriminate between them.

Instead of holding constant the high payoff and fixing the ratio of winning probabilities at 0.8 and varying the size of the intermediate payoff, in this treatment we held constant both the intermediate and high payoffs (at £20 and £40 respectively) and allowed the ratios of probabilities to vary. We presented one series of choices in which the certainty of £20 was compared with (£40,  $p$ ) with  $p$  varying from 0.1 to 0.9 in increments of 0.1. Another series scaled all of these down by half: that is, (£20, 0.5) was compared with (£40,  $q$ ) with  $q$  varying from 0.05 to 0.45 in increments of 0.05. A third series involved probabilities scaled to a quarter of their original level, so that (£20, 0.25) was compared with (£40,  $r$ ) with  $r$  varying from 0.05 to 0.2 in increments of 0.05. Overall, then, there were 22 different pairs, shown in random order mixed in with other choices and presented in four different waves.

Such a design allows us to examine whether the standard direction of CRE pattern occurs; and if it does, it offers the possibility of seeing whether the pattern is consistent with what we might expect from the kind of weighting function proposed in Tversky and Kahneman’s (1992) Cumulative Prospect Theory (CPT) or whether it seems more compatible with a similarity-based explanation.

We start with CPT. There are four ‘common ratios’ of probabilities that we can track through the three degrees of scaling. However, hardly any participants chose the risky lottery at any level of scaling when the ratio was 0.2:1, so we focus on the three common ratios as shown in Table 1.

The key idea in CPT is that the overall subjective value of any lottery is the weighted average of the subjective values  $v(\cdot)$  of each payoff, with the weights being some nonlinear transformation of the stated probabilities of receiving those payoffs. The columns in Table 1 show the ratio of decision weights when the probabilities are transformed according to the weighting function estimated in Tversky and Kahneman (1992).<sup>10</sup> According to CPT applied to the present choices, the value of the safer option will be greater than, equal to, or less than the value of the riskier option if and only if

<sup>10</sup> We report the figures from that particular function where the key coefficient  $\gamma$  was estimated to be 0.61, but the same general point holds for other estimates of  $\gamma$  in the range between 0.5 and 0.7, which is the range most commonly reported / asserted.

**Table 1** Ratios of decision weights for various common ratios

	Common ratio=0.8	Common ratio=0.6	Common ratio=0.4
£20, 1	$w(0.8)/w(1)=0.607$	$w(0.6)/w(1)=0.474$	$w(0.4)/w(1)=0.370$
£20, 0.5	$w(0.4)/w(0.5)=0.880$	$w(0.3)/w(0.5)=0.757$	$w(0.2)/w(0.5)=0.620$
£20, 0.25	$w(0.2)/w(0.25)=0.897$	$w(0.15)/w(0.25)=0.780$	$w(0.1)/w(0.25)=0.641$

the ratio of the subjective values of the intermediate and high payoffs—i.e.,  $v(£20)/v(£40)$ —is greater than the ratio of the weights as shown in any particular cell in Table 1. For example, when an individual is presented with the choice between (£20, 1) and (£40, 0.6), she will be indifferent between the two (or, when put probabilistically, will be stochastically indifferent between the two) when her  $v(£20)/v(£40)=0.474$ ; hence she will be more likely to choose the certainty if her  $v(£20)/v(£40)>0.474$  and more likely to choose the riskier option if her  $v(£20)/v(£40)<0.474$ .

What Table 1 shows for all three common ratios is that if there is to be any substantial increase in the likelihood of choosing the riskier option, it will come when the probabilities are scaled down by the ‘first’ half—i.e., when going from the pairs involving (£20, 1) to those involving (£20, 0.5)—since there is a considerable increase in the ratios of weights over this range in all three columns. For example, anyone whose  $v(£20)/v(£40)$  lies between 0.474 and 0.757 would move from having a greater than 0.5 probability of choosing the safer option in the {(£20, 1), (£40, 0.6)} pair to having a greater than 0.5 probability of choosing the riskier option in the {(£20, 0.5), (£40, 0.3)} pair. However, within each column the ratio of weights changes much less when scaling down from pairs involving (£20, 0.5) to those involving (£20, 0.25). Thus there is likely to be much less of an increase in the probability of choosing the riskier option over this range of scaling down: only those whose  $v(£20)/v(£40)$  lies between 0.757 and 0.780 would move from having a greater than 0.5 probability of choosing the safer option in the half-scaled case to a greater than 0.5 probability of choosing the riskier option in the quarter-scaled case. Much the same applies (albeit at different levels) to the other common ratios in Table 1.

By contrast, the similarity account offered by Rubinstein (1988) might be expected to give a different pattern. That account can be summarised as follows. When the alternatives are sufficiently different on both the payoff and the probability dimension, many people act more or less like risk averse EU maximisers, choosing the safer option as long as its expected value is not too much lower than the expected value of the riskier option. However, as the probabilities are progressively scaled down and become more and more similar, a point is reached when the difference between those probabilities is judged ‘inconsequential’, so that the probability dimension is ignored and the payoff advantage of the riskier option now becomes decisive. Although the model does not specify where this switching threshold occurs, it suggests that there is more likelihood of similarity-based action as the scaling factor goes from 0.5 to 0.25 than when it goes from 1 to 0.5.

So the CREPROBS questions were intended to explore behaviour that might shed some light upon the respective claims of the ‘probability-transformation’ and

‘similarity’ accounts and/or might stimulate designs for further attempts to discriminate between probabilistic versions of these or other non-EU core models.

## 4 Implementation and results

### 4.1 Implementation

We recruited 148 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 along 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. This number was automatically copied as a password to the experimental interface. This insured 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) online platform—see footnote 8 for links.

Each participant completed a total of 192 questions organised into five parts. One hundred eighty of the questions were some form of lottery-based task, mixed up and spread between parts 1, 3, and 5. Parts 2 and 4 consisted of quite different types of task. Each of these parts consisted of six hypothetical questions of the type used in the Domain Specific Risk Attitude (DoSpeRT) procedure (Blais and Weber 2006). In part 2 (part 4) we offered participants six activities which involved financial (health and safety) risks and asked them to indicate how likely they are to engage in each activity on a scale from one (highly unlikely) to seven (highly likely). At the end of the experiment, all participants were given individual feedback about their self-reported risk attitudes based on their answers in part 2 and part 4. These two parts created distractor tasks for our participants which allowed us to separate repetitions of the incentivised questions. Overall, 180 questions in parts 1, 3, and 5 were incentive-compatible while the 12 DoSpeRT questions in parts 2 and 4 were hypothetical.

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.<sup>11</sup> These randomly selected participants were invited to the DR@W experimental laboratory for individual scheduled appointments. One of the 180 lottery-based 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 amount. If she had chosen a lottery, she would see an opaque bag being filled with

---

<sup>11</sup> This experiment was in fact one of two being conducted in parallel using the same methodology. The other was concerned with examining the ‘preference reversal phenomenon’ and is described in a separate paper, Loomes and Pogrebn (2014), available on request. 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.

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.

The incentive mechanism was described to all participants in the instructions. Participants also received a practice question and had an opportunity to e-mail the experimental team in case they were not clear about the instructions. It took most participants 30–40 minutes to complete the online experiment. Each individual appointment in the DR@W laboratory lasted between 3 and 5 minutes. The average payoff in the experiment was approximately £12.

#### 4.2 Results

Perhaps the greatest concern about the data generated by this experiment is that four repetitions of each pair may be too few to obtain a really good fit for any curve. However, the majority of laboratory experiments with human participants will face practical constraints on the number of observations per individual and it is important to develop a method which can provide a good account of behaviour using a limited number of observations. So, rather than try to fit particular curves to each participant's sets of responses, we adopt a different approach.

Consider Fig. 7, which shows on the horizontal axis the different values of  $x_j$  in the CRESUMS questions, while the vertical axis measures the numbers of times B is chosen in any four repeated pairings with the same  $x_j$ .

The circles plot what we should expect to observe for an (illustrative) individual with deterministic preferences who is indifferent between B and some sure sum between £24 and £26: they will choose B on every occasion when  $x_j$  is less than or equal to £24 and never choose B when  $x_j$  offers £26 or more. We cannot identify the precise point of indifference, but we might proxy it as the midpoint—i.e., £25. In other words, when we count B being chosen a total of 20 times in this pattern, we take this to correspond with a best estimate of £25 for the point of indifference. Even if the truth is that the individual does not have deterministic preferences but his underlying probability of choosing B falls below one for some  $x_j$  greater than £24 and drops quite steeply to become zero for some  $x_j$  less than £26, it would still seem that the best we can do in the absence of any more detailed information is suppose that the best available estimate of his SI point is £25.

Were we to observe another seemingly deterministic individual choosing B just 16 times (i.e., always choosing B when  $x_j$  is less than or equal to £22 and never choosing B when  $x_j$  is £24 or more), our best estimate of the indifference value would be £23. Thus a difference of four in the B count corresponds with a difference of £2 in the SI point.

Now compare that with the (illustrative) individual whose responses are plotted as stars. She exhibits some variability for values of  $x_j$  between £22 and £30 inclusive: that is, for each of these pairs, she chooses the lottery on at least one occasion and also chooses the sure amount on at least one other occasion. For ease of exposition, the pattern shown here is symmetric around a best estimate of the SI point at  $x_j = £26$ . The B option is chosen a total of 22 times according to the stars in Fig. 7: that is, this individual chooses B twice more than the individual represented by the circles who had an indifference/SI point estimated at  $x_j = £25$ . So here too the number of times the fixed bet is chosen can act as the basis for estimating the SI point. Also, as a proxy for the

Total B  
Chosen

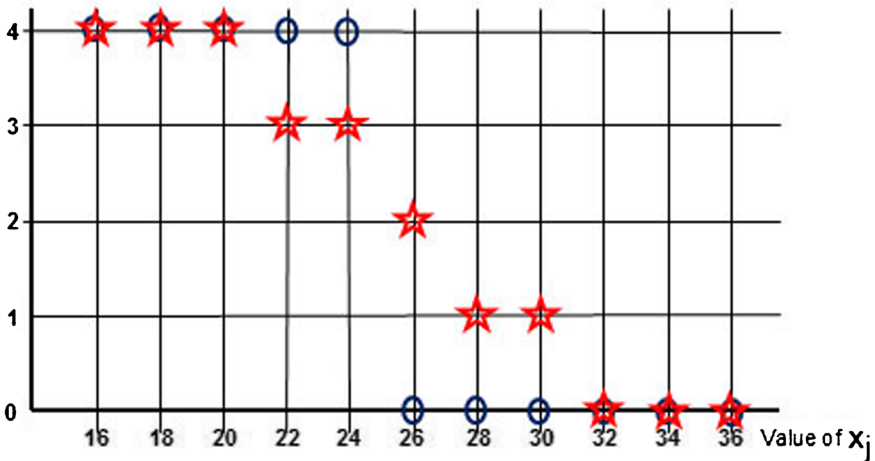


Fig. 7 Comparing deterministic and probabilistic responses

steepness/shalowness of the underlying probabilistic distribution, we can count the number of different  $x_j$  values for which some variability is observed: in Fig. 7, the deterministic individual scores 0 in this respect, whereas there are five such values (£22 to £30 inclusive) for the other individual. There will, of course, be some sampling error for such measures; but our sample sizes in conjunction with the within-subject nature of key parts of the analysis will still allow us to draw a number of conclusions.

We first consider the evidence concerning stochastic variability with respect to the 22 binary choice questions which are the focus of this paper. For each individual, we can count the number of pairs where some variability was exhibited: that is, where each alternative in that pair was chosen at least once in the course of the four repetitions. Table 2 shows how many of the total of 148 participants fell into each count category.

So out of 148 participants, the first column shows that there were 19 individuals who exhibited no variability: in these 22 pairs they chose the same option on all four occasions that each pair was presented to them. It is possible that at least some of these *might* have exhibited stochastic variability if the increments had been smaller than £2; but even if we take all 19 as having very precise step-function preferences, this only amounts to 13% of the sample. Another 11 only exhibited variability within a single pair, while 19 individuals made different choices within two pairs; and so on. The median and modal number of instances of within-person variability is four and the mean is 4.5 (reducing to 4.1 if we exclude five top-end outliers who accounted for 75 instances between them). So some degree of stochastic variability is quite typical.

**Table 2** Deterministic and stochastic choice patterns for 148 participants in the experiment

Pairs where each option was chosen at least once	0	1	2	3	4	5	6	7	8	9	10	>10
Numbers of Individuals	19	11	19	20	21	15	7	8	8	6	4	10



#### 4.2.1 Analysis of CRESUMS responses

Eighty-one individuals participated in this treatment. We first consider whether or not the standard CRE pattern is observed in those pairs which involved the classic parameters. To this end, we take the pair  $A=(£30, 1)$  and  $B=(£40, 0.8)$  and their scaled-down counterparts  $C=(£30, 0.25)$  and  $D=(£40, 0.2)$  and consider the various combinations for each of the four waves of binary choices, denoted by BC1 through BC4 to show the chronological order. Table 3 shows the results.

Although there was variability at the individual level, the aggregate patterns were quite stable across rows in the table and replicated the usual CRE pattern very clearly: although about half of all observations fell into the A&C and B&D columns and could be regarded as consistent with EUT, the other half were either A&D or B&C, with the asymmetry between these two being very pronounced—so much so, that in all four rows the aggregate modal preference switched from  $A > B$  to  $D > C$ .

This is a stronger effect than can be explained by a configuration such as the one shown in Fig. 3 where the SI is the same for  $\{A, B\}$  and  $\{C, D\}$  but where the slopes of the functions are different. If we take as a proxy for those slopes the counts of pairs where individuals choose each option on at least one out of the four presentations, we find a similar pattern for  $\{A, B\}$  and  $\{C, D\}$ , as shown in Table 4.

Analysis at the individual level shows no systematic difference in within-person variability between the two sets of 11 pairs: there is a positive and significant correlation between the frequency with which an individual exhibits variability for the  $\{A, B\}$  pairs and for the  $\{C, D\}$  pairs (Pearson  $r=0.467, p<0.01$ ; Spearman  $\rho=0.528, p<0.01$ ), consistent with the idea that, by and large, the slopes of an individual's underlying probability distribution of preferences between the scaled-up pairs are not very different from the slopes for the scaled-down pairs.

If this is the case, we should expect the CRE patterns in Table 3 to be explicable primarily in terms of shifts in the locations of the SI points for many of the participants. This is something we can examine up to a point by looking at the aggregate data. Table 5 shows the numbers of times (out of a possible  $81 \times 4 = 324$ ) that the riskier option was chosen from the scaled-up pair  $\{A, B\}$  and from the scaled-down pair  $\{C, D\}$  at each level of intermediate payoff. The conversion to a percentage is shown below each figure.

These figures represent the summation of all individuals' underlying preference distributions and therefore do not give the kind of breakdown provided in Table 3, but they do enable us to see that D is chosen more frequently than B at *all* values between

**Table 3** Patterns in the four waves of binary choices (BCs) when  $A = (£30, 1)$ ,  $B = (£40, 0.8)$ ,  $C = (£30, 0.25)$ ,  $D = (£40, 0.2)$

	A : B	C : D	A&C	B&D	A&D	B&C
1st wave (BC1)	59 : 22	21 : 60	20	21	39	1
2nd wave (BC2)	57 : 24	22 : 59	21	23	36	1
3rd wave (BC3)	61 : 20	17 : 64	15	18	46	2
4th wave (BC4)	56 : 25	20 : 61	19	24	37	1

**Table 4** Numbers of individuals exhibiting different levels of within-person variability

Pairs where each option was chosen at least once	0	1	2	3	4	5	6	7	8	9	10	11
{A, B}: Numbers of Individuals	24	16	9	14	6	4	2	4	1	1	–	–
{C, D}: Numbers of Individuals	29	14	10	9	6	2	3	4	1	–	2	1

£16 and £36. There is no sign here of the reverse-CRE, even when the riskier option is the modal choice in the scaled-up pair, as it is for all values of  $x_j$  equal to or less than £26.

To examine how far the patterns in Table 5 are due to shifts in different individuals’ underlying preference distributions, we proxy each individual’s SI point on the basis of counting the total number of times the fixed riskier lottery was chosen in all 44 of his/her binary choices across all 11 levels of  $x_j$ . On that basis, we categorise participants into three groups: those who chose D more often than B with a difference strictly greater than 4, representing a shift in SI point that would on average be equivalent to at least one £2 increment of  $x_j$ ; those for whom the difference in either direction was four or less, who we might (if we were being cautious) regard as having ‘near enough’ the same SI; and those who chose B more often than D with a difference strictly greater than 4. On this basis we have a breakdown of 55 : 22 : 4—which is to say that more than two-thirds of the sample exhibit repeated choice behaviour consistent with an SI value of  $x_j$  for the scaled-down pair which is more than £2 to the right of their SI value of  $x_j$  for the scaled-up pair.<sup>12</sup> In fact, for 39 of the 81 participants, the difference equates to £6 or more.

In short, these data strongly reject stochastic versions of EUT that seek to explain the CRE pattern simply in terms of a particular specification of the ‘random error’ term.<sup>13</sup> Rather, the results for the clear majority of participants look much more consistent with the kinds of patterns illustrated in Fig. 4.

However, if EUT does not provide an adequate core model, can the methods we have used help to discriminate between at least some of the non-EU alternatives? The other experimental treatment was designed to try to provide some guidance in that respect.

#### 4.2.2 Analysis of CREPROBS responses

Recall that the idea behind the design of this part of the experiment was to see whether, as the conventional shape of the CPT weighting function would entail, most of any switching from safer to riskier could be expected to occur when probabilities are initially scaled down by half, or whether, as similarity-based accounts might suggest, more switching could be expected when scaling down from a half to a quarter.

<sup>12</sup> Table 7 in the Appendix shows a more detailed breakdown of the differences which are here grouped into just three categories.

<sup>13</sup> The null hypothesis that there is no difference between the number of individuals with an SI point for B to the right of their SI point for D and the number of individuals with an SI point for B to the left of their SI point for D is rejected by a binomial test,  $p < 0.001$ .

**Table 5** Choices of B (D) versus alternatives offering various intermediate payoffs

Intermediate £	16	18	20	22	24	26	28	30	32	34	36
B Choices	256	247	216	205	184	167	157	91	51	44	21
%	79.0	76.2	66.7	63.3	56.8	51.5	48.5	28.1	15.7	13.6	6.5
D Choices	301	299	281	278	279	271	263	244	225	194	173
%	92.9	92.3	86.7	85.8	86.1	83.6	81.2	75.3	69.4	59.9	53.4

We keep the labels A and B for the fully scaled-up pairs, where in this treatment A stays fixed at (£20, 1) and B offers different probabilities of £40. As explained in Section 3.2, A was paired with nine different B lotteries, each repeated four times; but for reasons also explained in Section 3.2, we concentrate here on three levels of three common ratios and report the rest of the data in the Appendix. So our focus here will be upon the 12 choices where A was paired with  $B_1=(£40, 0.8)$ ,  $B_2=(£40, 0.6)$  and  $B_3=(£40, 0.4)$ , since these are the three  $B_j$ s that we can track at the different degrees of scaling discussed in relation to Table 1. We also keep the labels C and D for the quarter-scaled pairs. C was fixed at (£20, 0.25) and we consider the 12 choices involving C being paired four times with each of  $D_1=(£40, 0.2)$ ,  $D_2=(£40, 0.15)$  and  $D_3=(£40, 0.1)$ .

However, the CREPROBS treatment introduced an intermediate level of scaling down. So at the half-scaled level, we denote the fixed safer option by  $Y=(£20, 0.5)$ , while the three riskier lotteries of interest are denoted by  $Z_1=(£40, 0.4)$ ,  $Z_2=(£40, 0.3)$  and  $Z_3=(£40, 0.2)$ , with each  $\{Y, Z_j\}$  pair presented four times. To study the effect of going from fully scaled up to half scaled down, we can compare the number of times out of 12 that each individual chose A from  $\{A, B_j\}$  with the number of times the same individual chose Y from the 12 corresponding half-scaled-down  $\{Y, Z_j\}$  pairs.

Table 6 shows the stability of the aggregate patterns across all four waves BC1 through BC4. It also shows that when we compare the responses to the fully scaled-up  $\{A, B_j\}$  pairs with the responses to the quarter scaled-down  $\{C, D\}$  pairs, the usual shift from safer to riskier options is observed, especially for the probability ratios 1 : 0.8 and 1 : 0.6.

However, the data in Table 6 suggest that these switching patterns were not primarily due to underweighting higher probabilities, as the standard form of CPT would entail. In all 12 rows of the table, moving from  $\{A, B_j\}$  to  $\{Y, Z_j\}$  never once resulted in an increase in the numbers of riskier options chosen: in two rows the numbers stayed the same and in the other ten there was actually an increase in the numbers of safer options chosen.<sup>14</sup> By contrast, when comparing  $\{Y, Z_j\}$  with  $\{C, D_j\}$ , the increase in the numbers of risky options chosen is clear in all 12 rows and most pronounced in the top eight where the options were more evenly balanced. While this is not easy to reconcile with the standard CPT weighting function, it could be compatible with a similarity account.

While Table 6 is useful in showing the aggregate patterns, our data also allow us to examine shifts at the level of the individual participant. In each panel of Fig. 8, negative

<sup>14</sup> In the Appendix, Table 8 shows the numbers and corresponding percentages of riskier choices for all pairs, allowing us to see that the same tendency for more safe choices in the first phase of scaling down was also apparent for the cases where 0.9 was scaled down to 0.45, where 0.7 was scaled down to 0.35, and where 0.5 was scaled down to 0.25. In short, it was a very consistent tendency.

**Table 6** Aggregate patterns across repetitions and different degrees of scaling

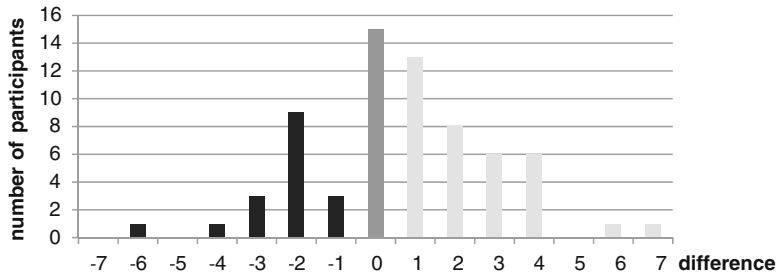
	{A, B <sub>1</sub> }	{Y, Z <sub>1</sub> }	{C, D <sub>1</sub> }
	<i>l</i> : 0.8	0.5 : 0.4	0.25 : 0.2
1st wave (BC1)	21 : 46	27 : 40	9 : 58
2nd wave (BC2)	20 : 47	24 : 43	13 : 54
3rd wave (BC3)	19 : 48	23 : 44	13 : 54
4th wave (BC4)	21 : 46	27 : 40	18 : 49
	{A, B <sub>2</sub> }	{Y, Z <sub>2</sub> }	{C, D <sub>2</sub> }
	<i>l</i> : 0.6	0.5 : 0.3	0.25 : 0.15
1st wave (BC1)	51 : 16	53 : 14	36 : 31
2nd wave (BC2)	44 : 23	53 : 14	41 : 26
3rd wave (BC3)	53 : 14	53 : 14	36 : 31
4th wave (BC4)	52 : 15	55 : 12	42 : 25
	{A, B <sub>3</sub> }	{Y, Z <sub>3</sub> }	{C, D <sub>3</sub> }
	<i>l</i> : 0.4	0.5 : 0.2	0.25 : 0.1
1st wave (BC1)	64 : 3	64 : 3	58 : 9
2nd wave (BC2)	64 : 3	67 : 0	61 : 6
3rd wave (BC3)	63 : 4	66 : 1	58 : 9
4th wave (BC4)	63 : 4	67 : 0	64 : 3

differences represent choosing the riskier option more often in any set of 12 choices as the probabilities are scaled down: that is, they represent movements in the ‘usual’ CRE direction. Thus the black bars show the numbers of individuals exhibiting a particular difference in the usual direction. Positive differences representing behaviour in the reverse-CRE direction are shown in light grey.

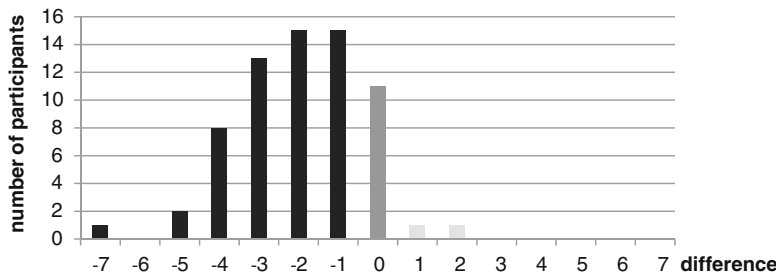
Going from {A, B<sub>j</sub>} to {Y, Z<sub>j</sub>} is the change that should, according to the conventionally shaped CPT weighting function, produce most reversals in the standard CRE direction. But on the basis of Fig. 8(a), we can see that substantially more individuals actually exhibited the reverse-CRE for this change: overall, 35 had a positive difference consistent with reverse-CRE while 17 had negative differences consistent with standard CRE and 15 exhibited the zero difference consistent with core independence. If our null hypothesis were that those who do not conform with independence are equally likely to exhibit the standard CRE pattern as the reverse-CRE pattern, we should reject that hypothesis at the 5% level in favour of a tendency towards reverse-CRE. If we wished to allow a difference of ±1 to be ‘near enough’ to EUT, then the breakdown between reverse-CRE, independence and standard CRE would be 22 : 31 : 14. In short, for the range of scaling down where the conventional CPT weighting function would entail the most movement in the standard CRE direction, we actually find some asymmetry in the opposite direction.

However, panel (b) in Fig. 8 shows a very different picture. When we compare half scaled-down pairs with quarter scaled-down pairs, 54 individuals registered differences in the usual CRE direction, while just 2 exhibited any reverse-CRE movement. Even if we regard a difference of ±1 as ‘near enough’ to EUT, the ratio of standard-CRE to reverse-CRE is still 39:1. It is in this panel that the really strong effects are seen—

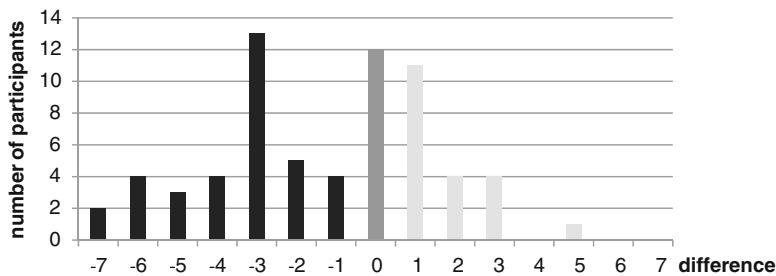
**a Comparison between {A, B<sub>j</sub>} and {Y, Z<sub>j</sub>}: fully scaled-up vs half scaled-down**



**b Comparison between {Y, Z<sub>j</sub>} and {C, D<sub>j</sub>}: half scaled-down vs quarter scaled-down**



**c Comparison between {A, B<sub>j</sub>} and {C, D<sub>j</sub>}: scaled-up vs quarter scaled-down**



**Fig. 8** CRE and reverse-CRE patterns for different degrees of scaling down

effects which are much easier to account for in terms of similarity judgments than in terms of the usual CPT configuration of probability weighting.

When the panel (a) and panel (b) trends are combined to provide a direct comparison between the scaled-up pairs and the corresponding quarter scaled-down pairs, we arrive at the pattern in panel (c). Here, 35 individuals choose the riskier option from {C, D<sub>j</sub>} more often than they choose the riskier option from {A, B<sub>j</sub>}, while 20 go in the opposite direction, with 12 exhibiting no difference.

So when we look at the types of questions often used in CRE experiments, involving comparisons between a fully scaled-up pair and a pair where the probabilities are scaled down by a quarter, we find the usual CRE pattern predominating in both of our treatments. However, when we break the scaling down into two parts, as in our

CREPROBS treatment, we find evidence that the overall pattern is not primarily due to the form of nonlinear transformations of probabilities suggested by CPT, but is more readily explained by similarity judgments. Of course, one should be wary of extrapolating too much from a single experimental result, but the strength and consistency of the patterns, within individuals and across repetitions, gives grounds for believing that there is at least a case for further investigation.

## 5 Concluding remarks

In this paper we have added to existing evidence about the stochastic variability in most people's preferences as expressed through binary choices. Some degree of such variability appears to be the norm—observable in at least 87% of our sample<sup>15</sup>—and although its extent varies from one individual to another, it appears to be prevalent and non-negligible.

We need to make allowance for this when trying to test decision theories formulated deterministically. As Fig. 2 showed, when we have the kind of step function entailed by deterministic models, the independence axiom has the same implications everywhere. So if C and D are, respectively, scaled-down versions of A and B, then *whenever* A is strictly preferred to B, C is strictly preferred to D, and vice-versa, with indifference between A and B requiring indifference between C and D. However, when preferences are probabilistic, the above relations only hold under strong restrictions. For weaker restrictions, the probability distributions relating to {A, B} and {C, D} are only required to correspond for the SI points.

This opens up the possibility that what might seem like systematic asymmetries at other points may (to some extent, at least) be produced by a particular choice of parameters. To provide a more comprehensive examination, we need to investigate a broader range of parameters, with sufficient repetition at the individual level to identify movements in the whole—or at least, the main part—of the probability distribution. That was the motivation for our experimental design.

In our experiment, the evidence strongly suggested systematic shifts in these distributions of a kind that are incompatible with EUT. However, we detected these shifts only when the degree of scaling down was sufficient to invoke similarity judgments. If CPT is to justify its current status as the front runner among alternatives to EUT, it should be able to organise the data from our CREPROBS treatment; but it cannot do so, except by radically altering the usual supposition about the shape of the probability transformation function.

We acknowledge that the coverage of our experiment is limited: one of the trade-offs involved in this kind of design is that many questions are required in order to get meaningful coverage of each probability distribution, so that each treatment can only focus on a few pairs. However, our results are by no means the only ones which suggest that, once we move away from the 'classic' CRE parameters, the robustness of the usual CPT assumptions is called into question.<sup>16</sup>

<sup>15</sup> In fact, if we also consider the other pairs used to separate and distract from the CRE pairs, the percentage of respondents exhibiting some variability during the session rises to 94%.

<sup>16</sup> See, for example, Battalio et al. (1990) and Prelec (1990), as well as cases of the 'reverse CRE' referred to earlier in the text. More recently, Harbaugh et al. (2010) found little support for CPT's predictions regarding risk attitudes in choice tasks.

But our main point is not to propose the abandonment of CPT in favour of some similarity-based model (or some other type of explanation)—although that is certainly a possibility that should be explored in future research. Rather, the main purposes of the paper have been (a) to argue for the importance of thinking in terms of probabilistic models and (b) to explore the possibility of an experimental design to investigate such models which is practicable within the constraints of working with human volunteers while also being capable of generating sufficient data to produce informative results at the individual level. We hope that extensions of—and no doubt, improvements upon—the ideas and designs in this paper will contribute to the development of richer descriptive models grounded in the probabilistic nature of human preferences.

**Acknowledgments** We gratefully acknowledge the financial support of the UK Economic and Social Research Council (grant no. RES-051-27-0248). Ganna Pogrebna thanks the Leverhulme Trust for financial support under the Leverhulme Early Career Fellowship scheme.

**Appendix**

**Table 7** CRESUMS: number of times D chosen minus number of times B chosen

Difference D–B	Number of individuals
≥ +36	11
+28 to +35	8
+20 to +27	5
+16 to +19	6
+12 to +15	9
+8 to +11	10
+5 to +7	6
+2 to +4	11
-1 to +1	9
-4 to -2	2
≤ -5	4

**Table 8** CREPROBS: aggregate data for all 22 pairs

A = (£20, 1)	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
B Choices	244	187	132	68	33	14	2	2	0
%	91.0	69.8	49.3	25.4	12.3	5.2	0.7	0.7	0
Y = (£20, 0.5)	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05
Z Choices	224	167	112	54	18	4	3	3	1
%	83.6	62.3	41.8	20.1	6.7	1.5	1.1	1.1	0.4
C = (£20, 0.25)		0.2		0.15		0.1		0.05	
D Choices		215		113		27		4	
%		80.2		42.2		10.1		1.5	

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

## References

- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton: Princeton University Press.
- Battalio, R., Kagel, J., & Jiranyakul, K. (1990). Testing between alternative models of choice under uncertainty: some initial results. *Journal of Risk and Uncertainty*, 3, 25–50.
- Bell, D. (1982). Regret in decision making under uncertainty. *Operations Research*, 30(5), 961–981.
- Bell, D. (1985). Disappointment in decision making under uncertainty. *Operations Research*, 33(1), 1–27.
- Blais, A.-R., & Weber, E. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 33–47.
- Blavatsky, P. R. (2010). Reverse common ratio effect. *Journal of Risk and Uncertainty*, 40(3), 219–241.
- Blavatsky, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: why both are important for analyzing decisions. *Journal of Applied Econometrics*, 25(6), 963–986.
- Butler, D. J., Isoni, A., & Loomes, G. (2012). Testing the ‘standard’ model of stochastic choice under risk. *Journal of Risk and Uncertainty*, 45, 191–213.
- Harbaugh, W., Krause, K., & Vesterlund, L. (2010). The fourfold pattern of risk attitudes in choice and pricing tasks. *Economic Journal*, 120, 595–611.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263–291.
- Leland, J. W. (1994). Generalized similarity judgments: an alternative explanation for choice anomalies. *Journal of Risk and Uncertainty*, 9(2), 151–172.
- Loomes, G. (2010). Modeling choice and valuation in decision experiments. *Psychological Review*, 117(3), 902–924.
- Loomes, G., & Pogrebna, G. (2014). Probabilistic preferences and failures of procedural invariance, mimeo.
- Loomes, G., & Sugden, R. (1982). Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal*, 92, 805–824.
- Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies*, 53(2), 271–282.
- Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. *Handbook of Mathematical Psychology*, 3, 249–410.
- Machina, M. (1982). Expected utility analysis without the independence axiom. *Econometrica*, 50(2), 277–323.
- Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, 59, 371–404.
- Prelec, D. (1990). A pseudo-endowment effect and its implications for some recent non-expected utility models. *Journal of Risk and Uncertainty*, 3, 247–259.
- Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4), 323–343.
- Rieskamp, J., Bussemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, 44, 631–661.
- Rubinstein, A. (1988). Similarity and decision-making under risk. *Journal of Economic Theory*, 46, 145–153.
- Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38(2), 332–382.
- Stott, H. (2006). Cumulative prospect theory’s functional menagerie. *Journal of Risk and Uncertainty*, 32(2), 101–130.
- Sugden, R. (2004). Alternatives to expected utility: Foundations. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory: Volume 2: Extensions* (pp. 685–755). Dordrecht: Kluwer.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.