



Prince: An improved method for measuring incentivized preferences

Cathleen Johnson¹ · Aurélien Baillon² · Han Bleichrodt² · Zhihua Li³ · Dennie van Dolder⁴ · Peter P. Wakker²

Accepted: 12 February 2021 / Published online: 31 July 2021
© The Author(s) 2021

Abstract

This paper introduces the Prince incentive system for measuring preferences. Prince combines the tractability of direct matching, allowing for the precise and direct elicitation of indifference values, with the clarity and validity of choice lists. It makes incentive compatibility completely transparent to subjects, avoiding the opaqueness of the Becker-DeGroot-Marschak mechanism. It can be used for adaptive experiments while avoiding any possibility of strategic behavior by subjects. To illustrate Prince's wide applicability, we investigate preference reversals, the discrepancy between willingness to pay and willingness to accept, and the major components of decision making under uncertainty: utilities, subjective beliefs, and ambiguity attitudes. Prince allows for measuring utility under risk and ambiguity in a tractable and incentive-compatible manner even if expected utility is violated. Our empirical findings support modern behavioral views, e.g., confirming the endowment effect and showing that utility is closer to linear than classically thought. In a comparative study, Prince gives better results than a classical implementation of the random incentive system.

Keywords Incentive compatibility · Random incentive system · BDM · Choice list · Matching

JEL Classification C91 · D81

✉ Peter P. Wakker
wakker@ese.eur.nl

¹ Economics Department, Chambers College of Business and Economics, West Virginia University, Morgantown, WV, USA

² Erasmus School of Economics, Erasmus University Rotterdam, 3000 DR Rotterdam, the Netherlands

³ Birmingham Business School, University of Birmingham, JG Smith Building, Birmingham B15 2TT, UK

⁴ School of Business and Economics, Free University Amsterdam, Amsterdam, the Netherlands

Behavioral economics challenges the classical revealed preference paradigm in economics. Many of the challenges could be handled by incorporating irrationalities and emotions in decision models, as for example in Tversky and Kahneman's (1992) prospect theory. However, preference reversals, revealing systematic differences between choice and matching,¹ entailed a more fundamental challenge. They cast doubt on the basic concept of preference. Although some researchers blamed choice-based procedures for preference reversals (Fischer et al. 1999), most researchers currently prefer choice to matching, following the recommendations by Arrow et al. (1993) and others. Binary choices have drawbacks too, though. They take more time to administer, give interval rather than point estimates, and have their own biases.² For this reason, some recent studies revealed choices from linear budget sets, an intermediate between binary choice and matching (Choi et al. 2007; Epper and Fehr-Duda 2015; Miao and Zhong 2015). This paper introduces a new incentive system to measure preferences that combines the greater validity of choice and the greater efficiency and precision of matching. It reconciles the two.

A pervasive difficulty in economic experiments is that real incentives as implemented in the laboratory are often difficult to understand for subjects. This problem is greatest for matching, where the Becker et al. (1964) mechanism (BDM) has often been criticized for this reason.³ Both choice and matching experiments commonly involve more than one decision. Paying at every decision leads to income effects. For this reason, the random incentive system (RIS; proposed by Savage 1954 p. 29) is now commonly employed. In this system, only one of the experimental decisions, randomly selected at the end, is implemented for real. If subjects treat each experimental decision as the only real one (isolation), incentive compatibility follows. However, subjects may conceive of the set of decisions as a meta-lottery (Holt 1986) where, for example, some decisions can be used to hedge others, resulting in possible spillover effects.

The Prince incentive method, introduced in this paper, reduces the aforementioned problems by combining and improving features from existing incentive systems, particularly the RIS, the BDM, and Bardsley's (2000) conditional information system. In brief, where capitalized letters explain the acronym Prince: (1) the choice question implemented for real is randomly selected PRior to the experiment; (2) subjects' answers are framed as INstructions to the experimenter about the real choice to be implemented at the end; (3) the real choice question is provided in a Concrete form, e.g., in a sealed envelope; (4) the Entire choice situation, rather than only one choice option, is described in that envelope. Incentive compatibility is now crystal clear, not only to homo economicus but also to *Homo sapiens*. In addition, isolation is maximally enhanced with the envelope with the one real choice situation in hand.

For adaptive (also called chained) experiments—where the sequence of questions is path dependent—strategic answering is impossible under Prince. Some recent adaptive

¹ In matching questions, subjects directly indicate indifference values. Attema and Brouwer (2013) provide a review.

² These biases have an older history in psychophysics (Gescheider 1997 Ch. 3). From the beginning (Fechner 1860), psychophysicists used binary comparisons besides matching to measure subjective values. The Nobel laureate von Békésy (1947) introduced bisection (the “staircase method”), to avoid the biases in choice lists (“limiting methods”). Williams (1966 p. 581) criticized choice lists for being crude.

³ See, for instance, Bardsley et al. (2010 p. 271–274), Harrison and Rutström (2008 §1.4, Appendix D), Holt and Smith (2016), and Seidl (2002 pp. 630–634).

experiments already made strategic answering practically impossible while still using the classical RIS incentive system. These experiments used computer algorithms to select the most informative stimulus to present to a subject given her previous choices (Cavagnaro et al. 2016; Chapman et al. 2018; Ryan et al. 2016; Toubia et al. 2013; Yang et al. 2020). Because the algorithms used to select future stimuli are too complicated for subjects to comprehend, it is practically impossible for them to increase their earnings by answering strategically. Unfortunately, the algorithmic complexity also implies that subjects cannot confirm for themselves that the experiment is indeed incentive compatible and must thus trust the experimenter on this issue. With Prince, incentive compatibility is crystal clear to subjects, and the impossibility to manipulate is so too for any subject who might think of it. We thus resolve the incentive compatibility problem for adaptive experiments, first raised by Harrison (1986), both practically and theoretically.

We apply Prince to incentivize Wakker and Deneffe's (1996) adaptive tradeoff method (TO) for measuring utility. This method provides parameter-free measurements of utility for expected utility that remain valid if expected utility is violated. Thus, it also retrieves the right utility function under prospect theory without requiring knowledge of the probability weighting function. However, implementing the method in an incentive compatible manner has not yet been possible. Using Prince, this is possible, so that the method becomes available to experimental economics. We can now measure risky utility in a parameter-free manner that, unlike other existing methods (Farquhar 1984; Holt and Laury 2002), is not distorted by the extensive violations of expected utility that have been found empirically. Finally, not only does our experiment avoid deception, but nondeception is transparently verifiable to subjects.

In addition to the adaptive experiment, we implement Prince in various standard experimental designs. These experiments are primarily intended to illustrate the wide applicability of Prince. Settling the classical questions in economics that they address will require more thorough investigations, which we leave for future research. The experiments do show the potential of Prince to contribute to these questions. All our findings support modern views: preference reversals signal a problem of our measurement instruments, the endowment effect is a genuine property of preference, ambiguity attitudes display aversion as well as insensitivity, and utility is closer to linear than traditionally thought (with risk aversion partly captured by components other than utility).

This paper is organized as follows. Section 1 defines Prince. The following sections describe the varied set of experiments implementing Prince, involving 251 subjects in total. For the sake of brevity, each experiment is described concisely in the main text, with details in the Electronic Supplementary Material. Section 2 implements Prince in a small, single-task experiment, showing how it combines the pros of matching and choice. Prince is also suited for multi-task experiments, as demonstrated in Section 3, and it solves the problem of manipulation in an adaptive utility measurement (Section 4). Section 5 compares Prince with a traditional RIS. Section 6 discusses preceding incentive systems that share components with Prince, such as using envelopes to specify choice situations prior to the experiment. A general discussion is in Section 7, followed by a summary of the pros and cons of Prince (Section 8). We conclude in Section 9. The Electronic Supplementary Material, §§ES.1-ES.15, provides all stimuli and additional analyses.

1 Prince explained

This section introduces the Prince system. We explain its principles in the first two subsections and define them formally in Section 1.3. Further discussion is in Sections 7 and 8.

1.1 Prince defined

The experiment begins with a *real choice situation* (RCS) randomly selected from a set of possible choice situations for each subject. In our experiments, the RCS is written on a slip of paper and put in a sealed envelope (following Bardsley 2000 p. 224). This envelope is given to subjects at the beginning of the experiment and remains sealed until the end. The RCS describes some choice options (two in our experiments; a mug versus a money amount in our first experiment). The subject will receive one of these options, and her goal in the experiment is to obtain the most preferred one.

Although the subject does not know the RCS in her envelope during the experiment, she does receive some information about the potential choice situations (e.g., average or range of outcomes employed) in advance. The partial description of the RCS is constructed so that each choice situation considered during the experiment may be the RCS. The subject does not need to know the exact probabilities of the latter possibility, and such probabilities do not need to be uniform. However, they should be salient enough to motivate subjects to answer the experimental questions truthfully (Bardsley et al. 2010 p. 220). It is important that the note in the selected prior envelope unambiguously describes the entire RCS, specifying *all* the choice options. Implementations of BDM in the literature that use prior envelopes commonly specify only one random prize in the envelope, not the entire choice situation. Other studies only specify a number in the envelope referring to the choice situation specified elsewhere, causing uncertainty for subjects about which RCS corresponds with their number. Further discussion and references are in Section 6.

During the experiment, various possible real choice situations are presented to the subject. We explicitly ask subjects to give “instructions” about the real choice to be implemented at the end of the experiment. This real choice is concrete, with the envelope held in hand. At the end of the experiment, the experimenter opens the envelope and selects the desired option based on the subject’s instructions. We never ask “what would you do if,” referring to nonconcrete choice situations. A script with statements such as “If you say what you want then you get what you want,” or “If you give wrong instructions, then you won’t get what you want” further emphasizes the connection between decision and outcome. This way, incentive compatibility is crystal clear to the subjects.

1.2 Prince for adaptive experiments: Problems and solutions

In adaptive experiments, stimuli depend on subject responses to previous stimuli. If traditional RISs are used, subjects may benefit (or think they benefit) from answering a question untruthfully to improve future stimuli. Such gaming is impossible with Prince, which is obvious to the subjects because the RCS—held in their hand—has been determined prior to the experiment.

In adaptive experiments, experimenters do not know exactly which choice situations will occur during the experiment. This raises two *overlap problems*.

- (1). The *indeterminacy overlap problem* arises if none of the subject's instructions relate to the RCS, leaving the choice from the RCS unspecified. The solution is simple: subjects may then choose on the spot.
- (2). The *exclusion overlap problem* arises if the partial information about the RCS excludes some choice situations generated during the experiment, thereby reducing salience and motivation for truthfulness in these excluded choice situations. To solve this problem, experimenters must frame this partial information about the RCS by anticipating the range of possible choice situations generated in the experiment. They can do this by using descriptive theories and pilots. For example, choice situations with very large monetary amounts could arise in our adaptive experiment (Section 4), depending on subjects' answers. We hence informed subjects about the existence of a large possible outcome ($> \text{€}3000$).

1.3 Prince summarized

We formally list the principles that define Prince for a given subject.

- (1). [PRIORITY] The RCS is determined at the start of the experiment *before the subject makes any decision*.
- (2). [INSTRUCTIONS TO EXPERIMENTER] We explicitly request "instructions" from the subject, asking what the experimenter should select from her RCS at the end, rather than asking vague "what would you prefer if" questions referring to unspecified situations.
- (3). [CONCRETENESS] The subject is given a description of the RCS in a tangible form, e.g., in a sealed envelope (the *prior envelope*), at the start of the experiment.
- (4). [ENTIRETY] The description handed to the subject completely and unambiguously describes the entire RCS.

For adaptive experiments, two criteria are added to the definition of Prince.

- (5). [NO INDETERMINACY] If none of the subject's instructions relate to the RCS, subjects can choose on the spot, after the envelope has been opened.
- (6). [NO EXCLUSION] The information about the RCS given to the subjects before the experiment should be framed so that it does not exclude potential choice situations faced during the experiment.

Although parts of Prince have been used before (Section 6), their integration into Prince is new. Their unifying principle is that they all make the subjects condition on the RCS. We expect that omitting any part will seriously weaken the effectiveness. For example, all BDM implementations that we are aware of violate Principle 4 [entirety]. This leads subjects to condition wrongly, enhancing rather than avoiding meta-lottery perceptions (see Section 7). This problem hampers BDM's internal validity and may account for its

bad performance. A detailed study of the effect of each component in Prince is left to future research.

In our experiments, all stimuli, including the prior envelopes and implementations of lotteries, are physical, but this is not essential for Prince. Other researchers may prefer computerized implementations. However, the physical availability of the RCS to every subject (e.g., in a prior envelope) is essential for Prince, which is why it is listed as Principle (3) above.

We avoid deception and, hence, the partial information about the RCS provided must be true. Although it is not a defining principle of Prince, subjects could completely verify the absence of deception in our implementations. They could always verify the correctness of the information provided about the stimuli. Second, unlike computer randomizations, our physically generated randomizations were also fully verifiable and were carried out by the subjects themselves.

2 Experiment 1: Reinventing matching (WTA)

Experiment 1 illustrates how to implement Prince in a small, single-task, experiment that uses a matching question to elicit one of the most used value concepts: willingness to accept (WTA). We measured WTA for a university mug that could be bought on campus for €5.95. WTA measures how much money a subject would accept in lieu of the mug. According to traditional theories, this should be the mug's cash equivalent.

$N = 30$ subjects (40% female) recruited from undergraduates from Erasmus School of Economics, Rotterdam, the Netherlands participated in one classroom session. Advertisement of the study promised a €10 show-up fee plus either a mug or additional money. Subjects immediately received a mug (*endowment*) along with the show-up fee.

Next, the experimenter presented 50 sealed envelopes, visibly numbered 1–50. These were separated into five piles of 10 each (1–10, ..., 41–50). Five subjects each checked one pile to verify that each number between 1 and 50 occurred once. Section ES.4.4 explains how subjects could completely verify that there was no deception. The subjects placed the envelopes into a large opaque bag, shuffled them, and randomly redistributed them over smaller bags (one for each row in the classroom). In turn, each subject randomly took one envelope, the prior envelope, from a bag (without replacement). Subjects were told that the note in their envelope described two options and that we would give them one of these two at the end of the session, based on their instructions to us.

Subjects received a questionnaire reproduced in Fig. 1, and were given a short written explanation and a PowerPoint presentation on the procedure. They were told that they could give up the mug for a price: “You will write instructions, for each possible content of your envelope (for each money amount), which of the two options you want. At the end, we will give you what you instructed. ... If you write what you want, then you get what you want!” The question in Fig. 1 is called Question 1 for later comparisons with Experiment 2.

At the end of the experiment, subjects handed in their questionnaire (instructions to experimenter). An experimenter opened their envelope, observed the real choice situation specified in the envelope, and followed the instructions in the questionnaire. The average WTA was 4.99 (SD 2.41). Further results are in Section 3.3.

In each of the 50 envelopes, one option is to keep your mug, and the other option is to give up your mug for a money amount. The note in each envelope is as follows.

Option 1: Keep your mug

Option 2: Give up your mug for € x

The money amount x varies between €0 and €10 in different envelopes. Five of the envelopes contain a randomly generated amount between €0 and €1, five envelopes contain a randomly generated amount between €1 and €2, five contain a randomly generated amount between €2 and €3, and so on, with finally five envelopes containing a randomly generated amount between €9 and €10. Thus the amount in your envelope can be any amount, in cents, between €0 and €10.

Please give us instructions, for each possible envelope that your envelope may be, whether we should let you keep your mug, or we should give you that money amount in exchange for your mug. Do so by specifying a threshold (in cents).

My threshold is €.....

If the money amount x in my envelope is equal to or above the threshold, then give me that money amount in exchange for my mug.
 If the money amount x in my envelope is below the threshold then let me keep my mug.

Fig. 1 Instructions for WTA with matching and endowment (Question 1)

Discussion of providing range 0–10 for answers Whereas specifying a range cannot be avoided for choice lists, it is optional for matching. We chose to specify it here, but for the sake of comparison will not do so later in Questions 5 and 6 in Section 3.5. There are pros and cons either way (Birnbau 1992). We chose the specific range here to facilitate comparability with choice lists presented later.

3 Experiment 2: Prince implemented in a multi-task experiment

Experiment 2 contrasts with Experiment 1 in showing how Prince can be used when there are many measurements, even though Prince implements only one choice for real.

3.1 General procedure

$N = 80$ subjects (41.2% female) recruited from undergraduates from Erasmus School of Economics, Rotterdam, the Netherlands, were randomly divided into two groups. Each group participated in one classroom session. They received a €10 show-up fee and

could gain an additional offering: money, mug, or chocolate. Procedural instructions, including a short presentation, were given by the experimenter (§ES.7). For each of the two sessions, there were 90 envelopes numbered 1–90 in random order. As in Experiment 1, these envelopes were separated into piles of ten, checked by subjects, shuffled, and randomly distributed without replacement.

The two groups of subjects received different versions of the first question, *1-match* or *1-choice* (see Figs. 2 and 3), which were variants of the question used in Experiment 1. These questions were part of a between-subject test in this larger experiment. To facilitate comparison with Experiment 1, these questions were always asked at the start of Experiment 2 (prior to the other questions in this experiment). The remaining eight questions, 2–9, were asked in random orders to all subjects in the two groups.

Instructions for envelopes of type γ

In each of the 10 envelopes of type γ , one option is the mug you just saw, and the other option is a money amount. The note in each envelope of type γ is as follows.

Type γ

Option 1: The mug

Option 2: €x

The money amount x varies between €0 and €10 in different envelopes. One of the envelopes contains a randomly generated amount between €0 and €1, one envelope contains a randomly generated amount between €1 and €2, one contains a randomly generated amount between €2 and €3, and so on, with finally one envelope containing a randomly generated amount between €9 and €10. Thus the amount in your envelope can be any amount, in cents, between €0 and €10.

Please give us instructions, for each possible envelope of type γ that your envelope may be, whether we should give you the money amount or the mug. Do so by specifying a threshold (in cents).

My threshold is €.....,.....

If the money amount x in my envelope is equal to or above the threshold, then give me that money amount.
 If the money amount x in my envelope is below the threshold then give me the mug.

Fig. 2 Instructions for cash equivalent with matching and no endowment (Question 1-match)

Instructions for envelopes of type δ

In each of the 10 envelopes of type δ , one option is the mug you just saw, and the other option is a money amount. The money amount x varies between €0.50 and €9.50 in different envelopes (see below).

The note in each envelope of type δ is as follows.

Type δ

Option 1: The mug

Option 2: € x

In the following list, each line describes the content of one envelope of type δ . On each line, cross out the square before the option that we should give you if that line describes the two options in your envelope.

1. <input type="checkbox"/> MUG	<input type="checkbox"/> €0.50
2. <input type="checkbox"/> MUG	<input type="checkbox"/> €1.50
3. <input type="checkbox"/> MUG	<input type="checkbox"/> €2.50
4. <input type="checkbox"/> MUG	<input type="checkbox"/> €3.50
5. <input type="checkbox"/> MUG	<input type="checkbox"/> €4.50
6. <input type="checkbox"/> MUG	<input type="checkbox"/> €5.50
7. <input type="checkbox"/> MUG	<input type="checkbox"/> €6.50
8. <input type="checkbox"/> MUG	<input type="checkbox"/> €7.50
9. <input type="checkbox"/> MUG	<input type="checkbox"/> €8.50
10. <input type="checkbox"/> MUG	<input type="checkbox"/> €9.50

Fig. 3 Instructions for cash equivalent with choice list and no endowment (Question 1-choice)

Each of the nine questions corresponded to a type (the term used with subjects) of envelope, and there were ten envelopes of each type. The numbering (1-match/choice, 2,3,...,9) of types/questions used in this paper was not communicated to subjects. Thus, each subject randomly drew an envelope, their prior envelope containing their RCS, from the 90 envelopes and then gave nine instructions in response to nine types/questions.

At the end of the experiment, subjects handed in their questionnaire and prior envelope. An experimenter opened the envelope, searched for the instruction in the questionnaire pertaining to the RCS, and carried it out.

Table 1 Statistics for Questions 1 (matching with endowment), 1a (matching without endowment), and 1b (choice list without endowment)

	Groups	N	Mean	SD
Experiment 1	Question 1	30	4.99	2.41
Experiment 2	Question 1-match	41	3.19	1.96
	Question 1-choice	38	3.61	2.54

3.2 The endowment effect

Question 1-match (Fig. 2) measured subjects' WTA for a mug, however, now without endowment. It was asked to 41 of the 80 subjects. Results and discussion are at the end of Section 3.3.

3.3 Matching versus choice lists between subjects

Question 1-choice (Fig. 3) repeats Question 1-match, again with no endowment, but now using choice lists instead of matching for the remaining 39 subjects. The sure amount of money (the alternative to the mug) increases with each option presented. At first, nearly all subjects preferred the mug, but nearly all subjects preferred the money by the end. At some point, they switched. We took the midpoint between the two money amounts where they switched as their indifference point.

An inconsistency results if a subject takes the money when the money offer is small but then switches to the mug when more money is offered. We allowed such inconsistencies so as to detect subjects' misunderstandings. The number of misunderstandings provides information about the transparency of Prince.

Results of questions 1, 1-match, and 1-choice In the 119 choice lists presented in this experiment (39 subjects here and all 80 subjects in Section 3.4), there was only one inconsistency; i.e., only one switch in the wrong direction (by subject 59). In otherwise comparable studies, typically 10% of subjects have inconsistent switches (Holt and Laury 2002). Because this one subject exhibited other anomalies as well (violating stochastic dominance in a later question), we removed her from this analysis. Leaving her in would not alter our results. Table 1 reports summary statistics, and Table 2 reports tests.

Discussion Prince confirms the endowment effect because WTA with endowment exceeds WTA without endowment.⁴ This suggests that the endowment effect, rational or not, reflects a genuine property of preference (Brosnan et al. 2012; Korobkin 2003 p. 1244), and not a bias in measurement.

Prince shows no difference between choice and matching. Our matching questions are very similar to the choice questions, directly referring to the choice in the prior envelope held in hand. Accordingly, their equality is no surprise. Our contribution here is methodological: we made matching look like choice, combining the virtues of both.

⁴ For references, see Schmidt and Traub (2009).

Table 2 Tests of equality of means

Questions	Treatment	mean difference	t	df	p
1 vs 1-match	endowment or not	1.81	3.48	69	0.001
1-match vs 1-choice	matching versus choice	-0.42	-0.82	77	0.41

The test of choice versus matching presented here was between subjects. For its result, not rejecting the null, to be convincing, statistical power should be sufficient. The fact that we obtain a highly significant result for the endowment effect suggests that power is sufficient. Furthermore, we confirm our finding in a higher-powered within-subject test for all 80 subjects in Section 3.4.

3.4 Matching versus choice lists within subjects

Questions 2 and 3 replicate Questions 1-match and 1-choice with chocolate (price €6.25) instead of a mug. Chocolates and mugs were used by Kahneman, Knetsch, and Thaler (1990), and many follow-up studies. Here we follow suit. Questions 2 and 3 were asked to each subject, allowing within-subject comparisons. The stimuli are in §ES.1. The average cash equivalent was 3.31 for matching and 3.26 for the choice list ($t_{79} = 0.28$, $p = 0.78$), unable to reject the null hypothesis of equality.

3.5 Testing preference reversals

We used Prince to test the classical preference reversal of Lichtenstein and Slovic (1971). Details are in §ES.2. For Question 4, the choice question, we used Fig. 4.

Option 1 was $4_{0.97}0$ (receiving €4 with probability 0.97 and €0 otherwise), called the *P-bet* in the literature because the gain probability is high. Option 2 was $16_{0.3}0$, called the *\$-bet* because it has a high minimum possible gain (in dollars when receiving its name; Lichtenstein and Slovic 1971). We also measured their cash equivalents in Questions 5 and 6, using analogs of Fig. 2, but without ranges for amount x , writing only “The amount x varies among the envelopes.” Although this means that very little is known about x ’s randomness, this affects neither the compatibility nor the transparency of incentives.

Normal preference reversals (higher CE of the *\$-bet* but, paradoxically, choosing the *P-bet*) occurred for 11% of the subjects, and the opposite preference reversals (higher CE of the *P-bet* but choosing the *\$-bet*) occurred for 7% of the subjects. These percentages are not significantly different ($p = 0.37$) and are infrequent enough to be explained as random choice inconsistencies (Schmidt and Hey 2004). We find no evidence of genuine preference reversals. As regards not having specified a range for matching, we found more choice anomalies here than for Question 1, where we had specified a range. Our finding thus illustrates that providing context can reduce distortions (Birnbau 1992).

Our finding deviates from other studies of preference reversals, where normal preference reversals are found in large majorities (surveyed by Seidl 2002). It suggests

Instructions for envelopes of type θ

In the 10 envelopes of type θ , both options are probability-contingent money. The note in each envelope of type θ is as follows.

Type θ

Option 1: €4 with a probability of 97%

Option 2: €16 with a probability of 31%

Please give us instructions whether we should give you option 1 or option 2 if your envelope is of type θ .

Give me option

Fig. 4 Instructions for choice between lotteries (Question 4 of Experiment 2), the choice question)

that preference reversals reflect errors in measuring preferences (procedural variance) rather than genuine properties of preferences such as intransitivities (Tversky, Slovic, and Kahneman 1990). Prince restores consistency between choice and matching,⁵ thus resolving preference reversals.

3.6 Measuring subjective probabilities and ambiguity attitudes

The RIS has been especially criticized in the study of ambiguity attitudes (Bade 2015; Oechssler and Roomets 2014), the topic of this section. Using questions 7, 8, and 9, we replicate the measurements of subjective probabilities and ambiguity attitudes by Baillon and Bleichrodt (2015 Study 1). They used classical choice lists, whereas we use Prince and matching. Details are in §ES.3. We measured the probability p such that

$$10_{E0} \sim 10_p 0.$$

E denotes an event explained as an observation from the Dutch AEX stock index, and 10_{E0} means that the subject receives €10 if E happens, and nothing otherwise. $10_p 0$ means that the subject receives €10 with objective probability p . The probability p giving the preceding indifference is called the *matching probability* of event E , denoted $m(E)$, and has often been used to measure ambiguity attitudes (Viscusi and Magat 1992). We measured it for three events:

⁵ A close study of the stimuli used in this paper renders support to our claims. Compare, for instance, matching questions as in Fig. 1 with choice questions as in Fig. 4. The choice relevance—which by rational standards is the same—is clearly conveyed to the subjects in both figures.

$E = A$ (Question 7): The Dutch AEX stock index increases or decreases by no more than 0.5% during the experiment.

$E = B$ (Question 8): The Dutch AEX stock index increases by more than 0.5% during the experiment.

$E = A \cup B$ (Question 9): The AEX stock index does not decrease by more than 0.5% during the experiment.

Our presentation of questions was similar to Fig. 2, with option 1 being 10_{E0} and option 2 being 10_{p0} , requesting that a threshold for p (instead of x) be specified. Baillon and Bleichrodt (2015) showed how we can use these observations to analyze ambiguity attitudes, using a nonadditivity index $m(A) + m(B) - m(A \cup B)$ to capture ambiguity attitudes. We replicated all their findings. In particular, the nonadditivity index was mostly positive, rejecting expected utility and confirming a (ambiguity-generated likelihood)-insensitivity (Trautmann and van de Kuilen 2015). These properties are genuine properties of preferences and not artifacts of measurement. Hence, Prince did not remove them. Validity is confirmed because we found the same phenomena on subjective probabilities as other experimental studies. Here, as throughout, the advantage of Prince is that we obtained our results more quickly (using matching instead of choice) and more precisely (point estimate instead of interval estimate) than preceding papers.

4 Experiment 3: Prince implemented in an adaptive experiment; measuring utility

We use an adaptive method to measure utility and show how Prince can resolve incentive compatibility problems by ruling out strategic answering. Exact stimuli, instructions, and details are in §ES.7. We first piloted the following procedures in two sessions, each with about ten graduate students who had had considerable exposure to decision theory. After the pilot, they were tasked with criticizing the procedures, especially concerning possible deception by the experimenter or the subjects. They did not find weaknesses.⁶ These students, as well as colleagues in informal pilots, confirmed procedural transparency and the absence of biases.

We use Wakker and Deneffe's (1996) adaptive tradeoff (TO) method to measure utility. This method is robust to violations of expected utility and provides a correct utility function for most nonexpected utility theories. Abdellaoui (2000) used it to measure probability weighting, and Dimmock et al. (2018) used it in a representative sample of several thousand respondents in the American Life Panel (ALP). Implementations so far were not incentive compatible. Prince makes this method available to economists by allowing for proper incentivization.

4.1 The preferences to be elicited for the tradeoff method

We measure indifferences $r_{pg}^j \sim r_{p}^{j-1}G$, $j = 1, \dots, 4$ (Fig. 5, with the usual notation for lotteries). Superscripts indicate the sequence of outcomes r^j . The experimenter chooses

⁶ A humorous suggestion was: "pull the fire alarm just when you have to pay €3000."

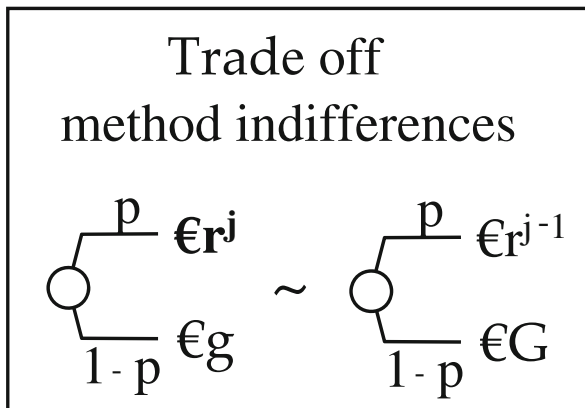


Fig. 5 Tradeoff method indifference

some preset values $0 < p < 1$, $0 < g < G$ (*gauge outcomes*), and $r^0 > G$. Then the bold-printed outcomes r^1 , r^2 , r^3 , r^4 are elicited sequentially from each subject over four stages. The experiment is adaptive because values r^j , after having been elicited, serve as input to the next question. We assume a *weighted utility* model:

$$\text{for } x \geq y, x_p y \text{ is evaluated by } \pi U(x) + \rho U(y) \quad (\pi > 0, \rho > 0). \quad (1)$$

This model includes expected utility, Quiggin's (1982) rank-dependent utility, Gul's (1991) disappointment aversion, prospect theory for gains (Tversky and Kahneman 1992), and most other generalizations of expected utility (Miyamoto 1988; Wakker 2010 §7.11). Eq. 1 implies that the r^j s are equally spaced in utility units (Wakker 2010 §3.3, §7.11, §10.6):

$$U(r^4) - U(r^3) = U(r^3) - U(r^2) = U(r^2) - U(r^1) = U(r^1) - U(r^0). \quad (2)$$

A nonparametric measurement of utility that is valid for most risky choice theories can be derived (Section 4.5, Section 4.6). Of course, the observations can also be used for parametric fitting (§ES.4.5 and §ES.4.6). The TO method avoids collinearity between utility U and probability weighting (π and ρ in Eq. 1): Eq. 2 is not affected by the probability weights π and ρ , and these do not even need to be estimated. For other measurements of prospect theory in the literature, collinearity is a serious problem (demonstrated by Zeisberger, Vrecko, and Langer 2012; p. 366–369). For a sophisticated measurement, see Bruhin, Fehr-Duda, and Epper (2010).

We carried out the TO measurement with four sets of predetermined values, one training set and three observational sets: TO0 (with t^j for r^j , t means training), TO1 (with x^j for r^j), TO2 (with y^j for r^j), and TO3 (with z^j for r^j) depicted in Fig. 6.

Wakker and Deneffe (1996) used the same stimuli but scaled up, and their choices were hypothetical.

Figure 7 displays the first two questions, TO1.1 and TO1.2, of the TO1 quadruple, as presented to the subjects. Question TO1.2 immediately followed TO1.1 on a separate page. Not only is the experiment adaptive, but it is also obviously so to subjects. Each

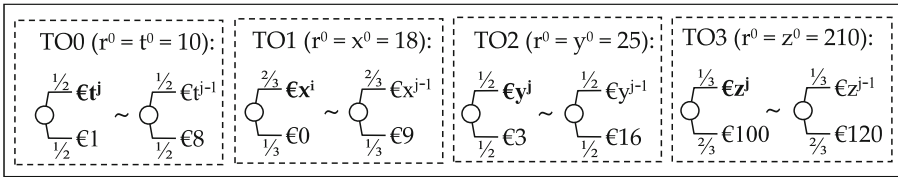


Fig. 6 The values used for TO0-TO3; $j = 1, \dots, 4$

subject had to impute the answer they gave to the first question ($x^1 = r^1$) before answering the next question (determining r^2). The third and fourth questions were like the second, requiring information of the previous answer.

4.2 Procedure and real incentives

We used Prince in a one-hour pen and paper session in a classroom. We conducted two sessions, one with 25 and one with 55 subjects. Subjects were undergraduate students from Erasmus University Rotterdam who were enrolled in an economics class. They

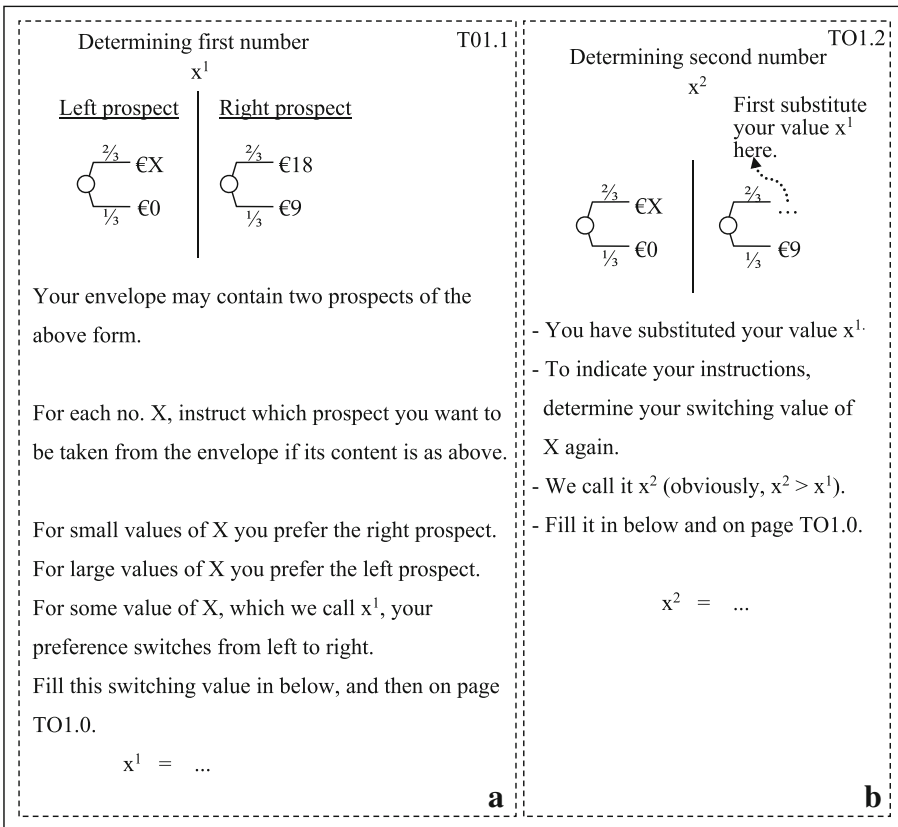


Fig. 7 Figures used in the tradeoff method

received a €5 show-up fee in addition to their performance-based payoff. They first chose a sealed envelope with their RCS. Then they received written explanations accompanied by an explanatory PowerPoint presentation.

Subjects filled out the training questions of TO0, jointly and simultaneously, exactly as in Wakker and Deneffe (1996), guided by the PowerPoint presentation. Subjects wrote their answers on pp. TO0.1–TO0.3, which they kept, but also on the front page TO0.0, which they tore off and gave to the experimenter at the end of the experiment. We explained how the performance payment procedure worked and how subjects' answers to the questionnaire determined the selection from the RCS in their envelope. Only then did subjects receive the three sets of questions TO1, TO2, TO3 (ordered randomly, subject-dependent), which they completed at their own pace. Three subjects in the first group and six in the second were randomly selected for real play. Their expected payoff if playing randomly (but subjects could, of course, do better) was €58.27. Under random play, the expected payoff over all 80 subjects totaled €10.99, in agreement with common policies of sufficient saliency of real incentives.

4.3 Construction and use of envelopes for real incentives, and avoiding the two overlap problems

In preparation for each session, we constructed 100 envelopes, from which each subject would randomly choose one (without replacement). Each envelope contained a slip of paper with two lotteries written on it (the RCS). We used popular theories of risky choice, mostly expected value and prospect theory, and pilot studies to determine the contents of the envelopes that minimize both overlap problems. The details depend on particularities of the experiment and are in §ES.4.2 and §ES.4.3.

4.4 Experiment with hypothetical choice

Besides the aforementioned sessions, we also conducted two sessions with hypothetical choices, one with 10 and one with 44 subjects. Subjects were unaware that other subjects played for real incentives. There was, of course, no role for Prince here. Subjects received €10 for participation. They made less on average than the real incentive condition, but the session took less time. The results that follow relate to the incentivized sessions, unless stated otherwise. We only describe results from the hypothetical choice experiment when they differ from the incentivized experiment.

4.5 Analysis

As all methods, the TO method can be used for parametric analyses. An advantage is that it can also be used for nonparametric analyses, i.e., without a commitment to any family of, or any shape of, utility functions (Wakker 2010 §9.4.2). Because this is a novelty of the method, we report it here. Parametric analyses are in §ES.4.5 and §ES.4.6.

To develop a nonparametric test of concavity, note that for strictly concave utility we have (with $r = x, y, \text{ or } z$, respectively)

$$r^{i+2} - r^{i+1} > r^{i+1} - r^i \quad \text{for all } i, \quad (3)$$

and for strictly convex utility we have

$$r^{i+2} - r^{i+1} < r^{i+1} - r^i \quad \text{for all } i. \quad (4)$$

We classified a subject's utility as concave if Eq. 3 was satisfied more often than Eq. 4, and as convex if the opposite held, with Eq. 4 satisfied more often than Eq. 3. The remaining subjects were irregular or linear.

4.6 Results

As regards the indeterminacy overlap problem of Section 1.2, for eight out of the nine envelopes opened during our experiment, the questionnaire answers determined the choice from the envelope, which was implemented. For the remaining, indeterminate, case the subject chose on the spot.

Figure 8 depicts the utility graphs resulting from average answers to the x , y , and z questions, based on Eq. 2, with normalizations $U(x^0) = 0$ and $U(x^4) = 1$, $U(y^0) = 0$ and $U(y^4) = 1$, and $U(z^0) = 0$ and $U(z^4) = 1$, respectively. These graphs do not involve parametric assumptions. They can also be produced for every individual. We can use overlaps of the x , y , and z regions to combine such curves into one overall curve on the union of domains.

As one would expect from the overall concavity of curves in Fig. 8, most subjects exhibit concave (Eq. 3) rather than convex (Eq. 4) utility: 37 versus 13 for the x 's, 29 versus 12 for the y 's, and 21 versus 14 for the z 's. This is significant for both x and y ($p \leq 0.01$), but not for z ($p = 0.31$). Our findings thus confirm moderate concavity of utility. For the x 's, the y 's, and the z 's, about 20% of our subjects had equality for all i in Eqs. 3 and 4, giving perfectly linear utility. The unclassified subjects exhibited irregular (or linear with noise) utilities. The hypothetical-choice groups' results were similar, but with more concavity for x and y stimuli, and not for z stimuli, than for incentivized groups.

We briefly summarize the results of our parametric analyses. For CRRA (constant relative risk aversion) utility, the median index of relative risk aversion was 0.04, and for CARA (constant absolute risk aversion) utility, the median risk tolerance was €10,000.

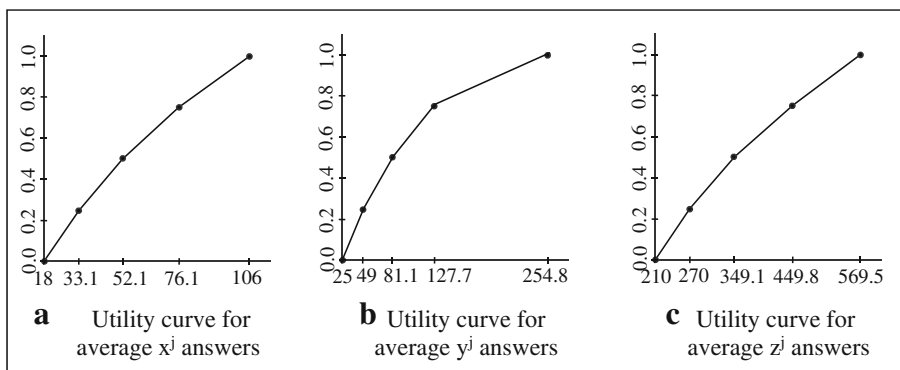


Fig. 8 Nonparametric utility curves for the TO method

CRRA fitted the data better than CARA. Both suggested weak concavity but did not deviate from linearity significantly. We found decreasing absolute risk aversion (Wakker 2010 p. 83) and decreasing relative risk aversion (Wakker 2010 p. 83 footnote 7).

Hypothetical data were noisier and contained more outliers. Further: (1) Hypothetical choice tended to have more risk seeking than real incentives in the z stimuli ($0.05 < p < 0.10$) both for CARA and CRRA utility. (2) No other significant differences were found between real and hypothetical choice.

4.7 Discussion of adaptive utility measurements

Cavagnaro et al. (2016), Chapman et al. (2018), and Toubia et al. (2013 p. 629) provide alternative suggestions for mitigating the problem of strategic answering in adaptive experiments. One of these suggestions, deriving a preference functional from the experimental answers and implementing this functional in the RCS, was implemented by Ding (2007). The downside of this approach is that subjects cannot directly understand the effects of their answers on the RCS during the experiment and have to trust the relevance of the derived functional. Relatedly, the computer-adapted stimuli used in these studies may be difficult for subjects to see through and manipulate, but in return, the exclusion overlap problem may be significant.

In experiments where subjects cannot really influence stimuli, they may mistakenly think they can, for example due to magical thinking (Rothbart and Snyder 1970) or illusions of control (Stefan and David 2013). Such distortions are more likely with future than with past uncertainties. Prince helps to avoid such distortions by determining the RCS before the subject makes actual decisions.

By classical economic standards, it may be surprising that we find near-linear utility, whereas classical estimates, based on expected utility, usually find more concavity. Recent studies find that risk aversion is mostly generated by factors other than utility for the moderate stakes considered in our experiment. With these factors filtered out, as in Eq. 2, utility turns out to be almost linear. Epper, Fehr-Duda, and Bruhin (2011), who like us, correct for deviations from expected utility, argue for the reasonableness of this finding.

We paid only a few randomly selected subjects and not all. An obvious drawback is that this makes the RCS less realistic. The advantage is that the stimuli can then involve large payments. This is desirable for utility measurement, where real curvature can only show up for large payments. This is why, after some initial debates, this procedure has now been widely accepted (Baltussen et al. 2012). Some authors have suggested that this system, with large payments for some subjects rather than moderate payments for all, improves subjects' motivation (Abdellaoui et al. 2011 Online Appendix A.2).

Unlike most measurements of utility in the literature, our analysis does not need to correct for deviations from expected utility. The TO stimuli were carefully devised so that those deviations have no bearing on our analysis, giving the same Eq. 2 under expected utility and nonexpected utility. The deviations are avoided rather than corrected for.

5 Experiment 4: Prince vs. the traditional random incentive system

A difficulty for preference theory, compared with many other empirical domains, is that there is no gold standard of true preference, as has often been pointed out (Bardsley

et al. 2010 Ch. 6; Infante, Lecouteux, and Sugden 2016; Pedroni et al. 2017 p.804 2nd para; Thaler and Sunstein 2008; Tversky and Kahneman 1981, 2nd to last para). Hence, there is no current consensus about the best methods for measuring preferences and no clear benchmark for a new method to beat. Prediction tests, commonly used to compare different models or theories for common data, usually cannot be used for different measurement methods because those involve different stimuli. Therefore, defenses of new (theories and) measurement methods are based primarily on internal validity arguments, using coherence criteria, stylized findings that the field has converged upon, and general psychological insights.⁷ Choi et al. (2007), Holt and Laury (2002), Andreoni and Sprenger (2012), and other introductions of new measurement methods shared this aspect with us. Therefore, they typically did not compare their method with an existing method.

Even so, we carry out an experimental comparison between Prince and a standard RIS, the most popular implementation of real incentives for individual preference today. Given the absence of a gold standard, our tests are mostly exploratory. We do speculate on true preferences for two tests, which we report here in the main text. Other tests are described only briefly here. All further experimental details and descriptions are in §ES.6.

We recruited 51 undergraduates from Erasmus School of Economics in Rotterdam. They were randomly divided into a group for Prince ($n = 26$) and a group for standard RIS ($n = 25$). They received €5 show-up fee and could gain an additional €11.50 on average from a lottery choice if they chose randomly.

For the first test, we replicated a violation of stochastic dominance (Tversky and Kahneman 1986). The choice is between two probability distributions (*lotteries*) over four outcomes depicted in Fig. 9, using obvious notation. The right lottery stochastically dominates the left one and should be preferred. We presented the choice twice to each subject, with many other choices in between so that subjects could be expected to have forgotten their previous choice and choose independently. Tversky and Kahneman (1986), and many replications (Birbaum and Navarrete 1998), found that most subjects prefer the left lottery. Subjects appear to ignore probabilities and erroneously think that there is an outcome-dominance. We claim that those majority preferences do not reflect true preferences. They result from misunderstanding due to the subtle framing of the complex lotteries, mainly due to ignoring probabilities. They will disappear if subjects fully understand the lotteries.

With RIS we replicated the majority preference (left). The average number of left choices per subject in the two choice situations was 1.42 (> 1 with $p = 0.02$; $n = 24$). However, with Prince, the average was exactly 1 ($p = 1$; $n = 26$), i.e., there were equally many choices for left as for right. Prince is closer to true preference (right) than RIS ($p = 0.04$) in this first test.

Our second test is a replication of Cox, Sadiraj, and Schmidt's (2015; CSS henceforth) comparative test. They considered five choices between a sure and a risky lottery and compared seven methods for implementing real incentives. CSS chose their One Task (OT) method as their gold standard for true preference. In OT, a subject carries out only one task that is implemented for real, avoiding income effects or meta-

⁷ Hence, we intentionally used stimuli from classic economic decisions throughout this paper, to illustrate the novelty and validity of Prince.

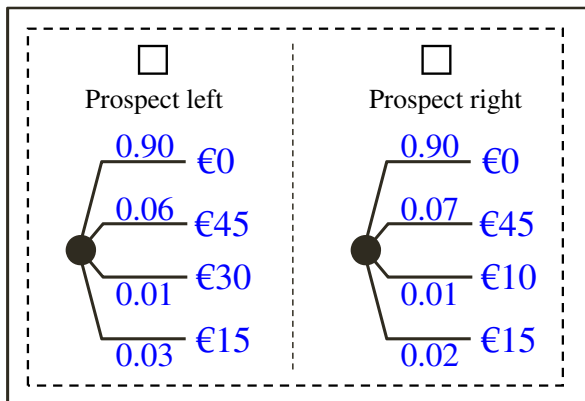


Fig. 9 Violation of stochastic dominance

lottery perceptions. Although this gold standard is not beyond debate (Binmore, Stewart, and Voorhoeve 2012 p. 234 point 4; Birnbaum 1992), experimental economists have commonly endorsed it (Bardsley et al. 2010 p. 268; Cox, Sadiraj, and Schmidt 2014, 2015; Starmer and Sugden 1991). We therefore adhere to it too and use CSS's OT results as a gold standard to have the same criterion throughout. It gives our measurements the extra handicap of between-experiment differences.

The non-italicized portion of Table 3 replicates CSS's Table 4, with the column "distance" added. The five columns S_1, \dots, S_5 correspond with their five lottery choices.⁸ We use these same lotteries. Rows describe implementation methods. We add our Prince and RIS to the bottom two rows. Cells in the table give percentages of safe choices. The last column gives the Euclidean distance to the gold standard (OT). The two best-performing methods of CSS are PAS (pay all, in a maximally correlated manner) and PAC/N (PAS, but divided by the number of questions, giving an average payoff). For CSS's other five methods we refer to their paper.

In all five choices, Prince is closer to the gold standard than RIS. Compared with the methods considered by CSS, Prince finishes third, defeated by their PAC/N and, closely, by their PAS. Prince's result is promising, the more so as it also faced the handicap of between-experiment differences. An advantage of Prince is that it is incentive compatible in revealing true preferences for homo economicus under common assumptions.⁹ PAS and PAC/N are subject to income effects and move preferences towards linear utility (Schmidt and Hewig 2015).

Our Prince-RIS comparison, Experiment 4, further reproduced several classical choice problems. We briefly summarize the results here. Prince tended to have lower risk aversion (marginally significant) and had significantly fewer common consequence violations of expected utility. It was closer to consistency and expected value

⁸ See Cox et al., (2015 Table 1): $3_{0.25}0$ vs $5_{0.20}0$, $10_{0.80}0$ vs $6_{0.25}0$ vs $10_{0.20}0$, $12_{0.75}6$ vs. $(0.75:12, 0.20:10, 0.05:0)$ (using obvious notation for three-outcome lottery), and $22_{0.80}12$ vs 18, respectively.

⁹ This theoretical claim does not need to assume expected utility, but only a weak dynamic isolation assumption (Bardsley et al. 2010 p. 269; Cohen et al. 1987), which Prince seeks to maximally enhance psychologically.

Table 3 Distance of methods to gold standard

Mechanism	S ₁	S ₂	S ₃	S ₄	S ₅	distance
OT (gold standard)	39.47	15.52	27.59	28.95	38.46	0
PORnp	37.50	45.00	47.50	32.50	60.00	41.78
PORpi	27.50	50.00	42.50	22.50	50.00	41.58
PORpas	22.50	42.50	20.00	10.00	30.00	38.78
PAS	25.64	23.08	33.33	10.26	17.95	32.42
PAC	36.84	52.63	23.68	21.05	42.11	38.41
PAC/N	37.50	35.00	35.00	22.50	45.00	22.86
PAI	36.84	52.63	36.84	34.21	52.63	41.21
RIS	12.00	40.00	08.00	60.00	68.00	59.79
Prince	23.08	38.46	23.08	34.62	53.85	32.93

maximization than RIS for all nonsignificant differences, regarding choice consistency, deliberate randomization, spillover effects, constant absolute risk aversion, constant relative risk aversion, and common ratio violations of expected utility. These findings may be interpreted as increased rationality, but their status vis-à-vis true preferences is unclear. We leave better calibrations of true preferences and more and better competitors for Prince to future studies.

6 Parts of Prince used in preceding studies

Virtually all preceding choice experiments using the RIS randomly select the RCS at the end of the experiment, thus violating our Principle 1 (priority) and Principle 3 (concreteness). All (to our knowledge) violate Principle 2 (instructions). Many satisfy Principle 4 (entirety), for example, randomly selecting a row in a choice list, which constitutes the entire choice situation. Virtually all matching experiments (mostly using BDM) violate Principles 1, 2, and 3, and none that we know of satisfy Principle 4. The remainder of this section focuses on studies (partly) satisfying Principle 1 by providing envelopes to subjects at the start of the experiment.

Loomes, Starmer, and Sugden's (1989) procedure in their experiments 1 and 2 comes close to Prince, also aiming to enhance isolation. The envelope selected a priori by each subject contained a number indicating the RCS, which concerned a choice between two lotteries specified later. This method violates priority because subjects cannot know whether the RCS corresponding with their number is determined a priori or only during the experiment, possibly depending on their answers. It violates Principles 3 and 4 because the envelopes do not contain the concrete RCS in its entirety, and subjects cannot know what RCS corresponds with their number. Adaptive manipulations by the experimenter, and therefore possibly by the subject, cannot be excluded. Epstein and Halevy (2018) similarly used a prior envelope containing the number of the RCS.

Bardsley (2000) partly satisfied Principle 1 too.¹⁰ Bardsley could not determine the choice options in the RCS for a given subject in advance because the latter depended on the choices made by other subjects during the experiment. Thus, Bardsley does not satisfy our Principles 2–4. He recommended Principle 3 (concreteness) for future studies (last paragraph of his §7).

In Schade, Kunreuther, and Koellinger (2012; first version 2001), options were determined a priori in an envelope (lying on a desk in the front of the room), but not whole choice situations. What was real (sculpture/painting) was determined only at the end of the experiment, and with a small probability. Hence, Principle 1 was partly satisfied, Principle 3 was approximately satisfied, but Principles 2 and 4 were not.

Bohnet et al. (2008) determined the RCS a priori. One choice option was inserted in an envelope visibly posted on a blackboard while subjects answered the experimental questions. Thus Principle 1 was satisfied, Principle 3 was approximately satisfied, but Principle 4 was not satisfied. The authors first asked subjects what they would “pick,” but later formulated these as instructions to the experimenters, thus partly satisfying Principle 2.

Studies using the Prince system of this paper include Baillon and Emirmahmutoglu (2018), Baucells and Villasís (2015), Bruhin, Santos-Pinto, and Staubli (2018), Calford (2020), Castillo (2020), and Li, Turmunkh, and Wakker (2019).

7 General discussion

The principles of Prince listed in Section 1.3 and, in general, all details of Prince serve to enhance isolation by enhancing psychological conditioning upon the RCS, increasing internal validity. Although Starmer and Sugden (1991)¹¹ found isolation satisfied in the RIS, violations have also been found. In fact, any finding of learning, order effect, or spillover effect (Baltussen et al. 2012; Cox, Sadiraj, and Schmidt 2014; Stewart, Reimers, and Harris 2015) in RIS implementations entails a violation of isolation, and such effects have been widely documented. Hence, improvements of isolation are desirable.

Regarding Principle 1 (priority), many studies have shown that conditioning works better for events determined in the past, even if yet uncertain, than for events to be determined in the future.¹² In the case of future determination, a meta-lottery is

¹⁰ The first experiment with a prior envelope may have occurred earlier, by Johann Wolfgang von Goethe (January 16, 1797, letter cited by Mandelkowitz 1968, p. 254). Goethe wrote: “I am inclined to offer Mr. Vieweg from Berlin an epic poem, Hermann and Dorothea ... Concerning the royalty we will proceed as follows: I will hand over to Mr. Counsel Böttiger a sealed note which contains my demand, and I wait for what Mr. Vieweg will suggest to offer for my work. If his offer is lower than my demand, then I take my note back, unopened, and the negotiation is broken. If, however, his offer is higher, then I will not ask for more than what is written in the note to be opened by Mr. Böttiger.” We thank Uyanga Turmunkh for this citation.

¹¹ They assumed one single choice per subject as gold standard for the implementation of real incentives, but Bimbaum (1992) criticized this assumption. This gold standard is also impractical for collecting rich data.

¹² See Bardsley et al. (2010 p. 277) and Shafir and Tversky (1992 p. 463). In Bardsley et al.’s (2010) terminology, Prince uses the direct decision approach and avoids the strategy method. Many studies (e.g., Halevy 2007; Kreps and Porteus 1979) have demonstrated that the timing of the resolution of uncertainty, even if of no strategic or informational relevance, still affects subjects and plays a role in time inconsistencies. In particular, prediction and postdiction are perceived differently (Heath and Tversky 1991 p. 9; Rothbart and Snyder 1970). Importantly for Prince, people more readily condition on uncertainties determined in the past than in the future, and take future uncertainty more as a meta-lottery (Keren 1991). This phenomenon underlies several findings in game theory (Weber, Camerer, and Knez 2004: virtual observability).

realistically perceived because the situation is still unresolved. More generally, we want the RCS to be perceived as realistically as possible. Advanced planning generates a psychological distance (Bardsley et al. 2010 §6.4.3). Strategy choice (subjects commit to all choices at the beginning of the experiment) further obstructs isolation by referring to random options (as with the random prizes of BDM) rather than to random choice situations. Principles 3 (concreteness) and 4 (entirety) reduce such obstructions.

There have been several implementations of real incentives using prior envelopes (Section 6) after Bardsley (2000), but all describe only one choice option in the envelope. If the randomization concerns the entire choice situation as with Prince (Principle 4), subjects can immediately condition on it, serving isolation. BDM randomizes a choice option (the price) rather than the choice situation, leading subjects to condition wrongly. It obfuscates the choice situation, with the random price draw enhancing the undesirable perception of meta-lotteries. Principle 4 (entirety) is crucial for Prince.

Researchers in decision theory will immediately see that Prince is strategically equivalent to RIS, soliciting real preferences. Homo economicus will behave the same in both procedures, and Prince need not be developed for her. However, as Bardsley et al. (2010 p. 270–271) wrote: “the effects of incentive mechanisms can depend on features of their implementation which are irrelevant from a conventional choice-theoretical point of view.” Prince seeks to minimize the biases generated by those features. It targets *Homo sapiens*.

Throughout the history of preference measurement, discussions about the pros and cons of matching versus choice have taken place.¹³ Choice is less precise. It takes more time to elicit preferences, requires a specification of range and initial values that generates biases, and enhances the use of qualitative noncompensatory heuristics (lexicographic choice and misperception of dominance). Matching is more difficult for subjects to understand, as are its incentive-compatible implementations. Further, the matching environment can lead subjects to ignore qualitative information and resort to inappropriate arithmetical operations.

Prince avoids an important misperception of matching: subjects may misperceive matching as bargaining.¹⁴ In Prince, with the choice situation (the price therein being one option) specified in advance in an envelope held in hand, it is perfectly obvious that this price is not subject to bargaining or any other influence.

Several experimental economists have implemented more than one choice situation for real, which is acceptable if the distortions due to the income effect are smaller than other distortions.¹⁵ Cox, Sadiraj, and Schmidt (2015) provide a systematic study, which is close in spirit to our study in seeking to reduce distortions in the RIS. It considers alternative incentive systems that imply particular income effects and investigates

¹³ These include Bostic et al. (1990), Hardisty et al. (2013), and Poulton (1989). There is also an extensive literature in the health domain (Stevens, McCabe, and Brazier 2007) and in psychophysics (Gescheider 1997 Ch. 3).

¹⁴ Because the link to the RCS is not clear in classical implementations, subjects think of what is closest to their everyday life, which is probably bargaining. See Korobkin (2003 p. 1243) and Sayman and Öncüler (2005 §2.2); also see Bardsley et al. (2010 p. 273).

¹⁵ Repeated payment is common in game and market experiments. It is not very common in individual choice, but still has been used in several studies, including Cox, Sadiraj, and Schmidt (2015), Epper, Fehr-Duda, and Bruhin (2011), and Mosteller and Nogee (1951). It will not generate large biases if linear utility of money can be assumed.

circumstances in which these income effects generate smaller distortions than the regular RIS does. Our study seeks to improve the RIS while avoiding any income effect, thus preserving incentive compatibility for homo economicus, rather than replacing the RIS by another system with some income effect.

This paper focuses on individual choice. The follow-up paper Li, Turmunkh, and Wakker (2019) adapts Prince to game theoretic experiments that include interactions between subjects. The envelopes for different players in a game were chosen jointly there.

The beginning of Electronic Supplementary Material, §ES.1, discusses an afterthought, suggesting an improved, more symmetric, formulation of the implementation payment in matching-type questions as in Figs. 1 and 2.

8 Pros and cons of Prince summarized

We have provided theoretical arguments for Prince showing (1) its internal validity (Section 1 and Section 7); (2) that it combines the pros of choice and matching, resolving a long-standing debate; (3) that it avoids the problem of strategic answering in adaptive experiments. We have also provided empirical arguments for Prince: (1) it induces highly consistent reporting (only one of 119 choice lists was inconsistent); (2) debriefings and discussions in pilots confirm its transparencies; (3) it confirms well-established preference findings; (4) it performs better than a standard RIS in a comparative study; (5) it reconciles choice and matching in four tests. Although such a reconciliation need not always be an improvement (Ariely, Loewenstein, and Prelec 2001), the preceding arguments suggest that it is.

Investigations of external validity are desirable. We can obtain useful insights into Prince's descriptive performance by investigating out-of-sample predictive power (especially regarding real-life decisions), extensive consistency checks to assess noise, and manipulations of Prince with separate principles turned on and off where Prince is compared with existing methods in these regards. This can reveal which component of Prince has which effect. The main purpose of this paper is to show that Prince as a whole works well and can be used to reduce documented violations of isolation. Given the length of this paper (and of the Supplementary Material), showing that Prince can be implemented for virtually every preference measurement, we prefer to leave the tests mentioned earlier to future studies. Contributions by objective outsiders will be especially useful.

Prince's main drawback is that it requires a nontrivial preparation by the experimenters: they must prepare envelopes with different choice situations for every session.

9 Conclusion

The Prince incentive system improves the standard random incentive system, the Becker-DeGroot-Marschak system, and Bardsley's (2000) conditional information system. Our subjects understand that there is only one real choice situation: the one they hold in hand. Prince resolves or reduces: (a) violations of isolation; (b) misperceptions of bargaining; (c) strategic answering in adaptive experiments. Incentive compatibility is completely transparent to subjects. Hence, we found virtually no irrational preference switch in choice lists.

An important contribution of Prince is that it revives matching. Prince makes it possible to combine the efficiency and precision of matching with the (improved) transparency and validity of choice. Prince avoids the major weakness of BDM by not randomizing choice options but instead whole choice situations, thus leading subjects to condition properly. Prince can serve to shed new light on classical questions in economics.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11166-021-09346-9>.

Acknowledgments Helpful discussions with Chen Li are gratefully acknowledged. Johnson thanks Social Philosophy and Policy Foundation for supporting her as a Research Scholar.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46(11), 1497–1512.
- Abdellaoui, M., Baillon, A., Placido, L., & Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review*, 101(2), 695–723.
- Andreoni, J., & Sprenger, C. (2012). Estimating time preference from convex budgets. *American Economic Review*, 102(7), 3333–3356.
- Ariely, D., Loewenstein, G. F., & Prelec, D. (2001). ‘Coherent arbitrariness’: Stable demand curves without stable preferences. *Quarterly Journal of Economics*, 118, 73–106.
- Arrow, K. J., Solow, R. M., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). Report of the NOAA panel on contingent valuation. *Federal Register*, 58, 4602–4614.
- Attema, A. E., & Brouwer, W. B. F. (2013). In search of a preferred preference elicitation method: A test of the internal consistency of choice and matching tasks. *Journal of Economic Psychology*, 39, 126–140.
- Bade, S. (2015). Randomization devices and the elicitation of ambiguity-averse preferences. *Journal of Economic Theory*, 159, 221–235.
- Baillon, A., & Bleichrodt, H. (2015). Testing ambiguity models through the measurement of probabilities for gains and losses. *American Economic Journal: Microeconomics*, 7, 77–100.
- Baillon, A., & Emirmahmutoglu, A. (2018). Zooming in on ambiguity attitudes. *International Economic Review*, 59(4), 2107–2131.
- Baltussen, G., Post, G. T., van den Assem, M. J., & Wakker, P. P. (2012). Random incentive systems in a dynamic choice experiment. *Experimental Economics*, 15(3), 418–443.
- Bardsley, N. (2000). Control without deception: Individual behavior in free-riding experiments revisited. *Experimental Economics*, 3(3), 215–240.
- Bardsley, N., Cubitt, R. P., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton: Princeton University Press.
- Baucells, M., & Villasís, A. (2015). Equal tails: A simple method to elicit utility under violations of expected utility. *Decision Analysis*, 12(4), 190–204.
- Becker, G. M., de Groot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3), 226–232.

- Binmore, K., Stewart, L., & Voorhoeve, A. (2012). How much ambiguity aversion? Finding indifferences between Ellsberg's risky and ambiguous bets. *Journal of Risk and Uncertainty*, 45(3), 215–238.
- Birnbaum, M. H. (1992). Should contextual effects in human judgment be avoided? Book review of: E. Christopher Poulton (1989), *Bias in Quantifying Judgments*, Hillsdale, NJ: Erlbaum. *Contemporary Psychology*, 37(1), 21–23.
- Birnbaum, M. H., & Navarrete, J. B. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, 17(1), 49–78.
- Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98(1), 294–310.
- Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior and Organization*, 13(2), 193–212.
- Brosnan, S. F., Jones, O. D., Gardner, M., Lambeth, S. P., & Schapiro, S. J. (2012). Evolution and the expression of biases: Situational value changes the endowment effect in chimpanzees. *Evolution and Human Behavior*, 33(4), 378–386.
- Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78, 1375–1412.
- Bruhin, A., Santos-Pinto, L., & Staubli, D. (2018). How do beliefs about skill affect risky decisions? *Journal of Economic Behavior and Organization*, 150, 350–371.
- Calford, E. M. (2020). Uncertainty aversion in game theory: Experimental evidence. *Journal of Economic Behavior and Organization*, 176, 720–734.
- Castillo, G. (2020). The attraction effect and its explanations. *Games and Economic Behavior*, 119, 123–147.
- Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, 52(3), 233–254.
- Chapman, J., Snowberg, E., Wang, S., & Camerer, C. (2018). Dynamically optimized sequential experimentation (DOSE) for estimating economic preference parameters. Working paper.
- Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97(5), 1921–1938.
- Cohen, M., Jaffray, J.-Y., & Said, T. (1987). Experimental comparisons of individual behavior under risk and under uncertainty for gains and for losses. *Organizational Behavior and Human Decision Processes*, 39(1), 1–22.
- Cox, J. C., Sadiraj, V., & Schmidt, U. (2014). Asymmetrically dominated choice problems, the isolation hypothesis and random incentive mechanisms. *PLoS One*, 9(3), e90742.
- Cox, J. C., Sadiraj, V., & Schmidt, U. (2015). Paradoxes and mechanisms for choice under risk. *Experimental Economics*, 18(2), 215–250.
- Dimmock, S. G., Kouwenberg, R., Mitchell, O. S., & Peijnenburg, K. (2018). Household portfolio underdiversification and probability weighting: Evidence from the field. NBER working paper series 24928, <http://www.nber.org/papers/w24928>.
- Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research*, 44(2), 214–223.
- Epper, T., & Fehr-Duda, H. (2015). Risk preferences are not time preferences: Balancing on a budget line: Comment (#12). *American Economic Review*, 105(7), 2261–2271.
- Epper, T., Fehr-Duda, H., & Bruhin, A. (2011). Viewing the future through a warped lens: Why uncertainty generates hyperbolic discounting. *Journal of Risk and Uncertainty*, 43, 163–203.
- Epstein, L. G., & Halevy, Y. (2018). Ambiguous correlation. *Review of Economic Studies*, 86, 668–693.
- Farquhar, P. H. (1984). Utility assessment methods. *Management Science*, 30(11), 1283–1300.
- Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig: Von Breitkopf und Härtel.
- Fischer, G. W., Carmon, Z., Ariely, D., & Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, 45(8), 1057–1075.
- Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Lawrence Erlbaum Associates.
- Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, 59(3), 667–686.
- Halevy, Y. (2007). Ellsberg revisited: An experimental study. *Econometrica*, 75(2), 503–536.
- Hardisty, D. J., Thompson, K. F., Krantz, D. H., & Weber, E. U. (2013). How to measure time preferences: An experimental comparison of three methods. *Judgment and Decision Making*, 8, 214–235.
- Harrison, G. W. (1986). An experimental test for risk aversion. *Economics Letters*, 21(1), 7–11.
- Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. In J. C. Cox & G. W. Harrison (Eds.), *Risk Aversion in Experiments*. Research in Experimental Economics Vol. 12. Bingley, UK: Emerald.

- Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4(1), 5–28.
- Holt, C. A. (1986). Preference reversals and the independence axiom. *American Economic Review*, 76, 508–513.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Holt, C. A., & Smith, A. M. (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics*, 8, 110–139.
- Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1), 1–25.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98(6), 1325–1348.
- Keren, G. B. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77(3), 217–273.
- Korobkin, R. (2003). The endowment effect and legal analysis. *Northwestern University Law Review*, 97, 1227–1293.
- Kreps, D. M., & Porteus, E. L. (1979). Dynamic choice theory and dynamic programming. *Econometrica*, 47(1), 91–100.
- Li, C., Turmunkh, U., & Wakker, P. P. (2019). Trust as a decision under ambiguity. *Experimental Economics*, 22(1), 51–75.
- Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89(1), 46–55.
- Loomes, G., Starmer, C., & Sugden, R. (1989). Preference reversal: Information-processing effect of rational non-transitive choice? *Economic Journal*, 99(395), *Supplement: Conference Papers*, 140–151.
- Miao, B., & Zhong, S. (2015). Risk preferences are not time preferences: Separating risk and time preference: Comment (#13). *American Economic Review*, 105(7), 2272–2286.
- Mandelkow, K. R. (1968; Ed). *Goethes Briefe*. Hamburg: Wegner.
- Miyamoto, J. M. (1988). Generic utility theory: Measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology*, 32(4), 357–404.
- Mosteller, F., & Noguee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, 59(5), 371–404.
- Oechssler, J., & Roomets, A. (2014). Unintended hedging in ambiguity experiments. *Economics Letters*, 122(2), 243–246.
- Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11), 803–809.
- Poulton, E. C. (1989). *Bias in quantifying judgments*. Hillsdale: Erlbaum.
- Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behaviour and Organization*, 3(4), 323–343.
- Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1), 128–154.
- Rothbart, M., & Snyder, M. (1970). Confidence in the prediction and postdiction of an uncertain outcome. *Canadian Journal of Behavioral Science*, 2(1), 38–43.
- Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley. (2nd edn. 1972, New York: Dover Publications).
- Sayman, S., & Öncüler, A. (2005). Effects of study design characteristics on the WTA-WTP disparity: A meta analytic framework. *Journal of Economic Psychology*, 26(2), 289–312.
- Schade, C., Kunreuther, H., & Koellinger, P. (2012). Protecting against low-probability disasters: The role of worry. *Journal of Behavioral Decision Making*, 25(5), 534–543.
- Schmidt, B., & Hewig, J. (2015). Paying out one or all trials: A behavioral economic evaluation of payment methods in a prototypical risky decision study. *Psychological Record*, 65(2), 245–250.
- Schmidt, U., & Hey, J. D. (2004). Are preference reversals errors? An experimental investigation. *Journal of Risk and Uncertainty*, 29(3), 207–218.
- Schmidt, U., & Traub, S. (2009). An experimental investigation of the disparity between WTA and WTP for lotteries. *Theory and Decision*, 66(3), 229–262.
- Seidl, C. (2002). Preference reversal. *Journal of Economic Surveys*, 16(5), 621–655.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24(4), 449–474.

- Starmer, C., & Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. *American Economic Review*, *81*, 971–978.
- Stefan, S., & David, D. (2013). Recent developments in the experimental investigation of the illusion of control. A meta-analytic review. *Journal of Applied Social Psychology*, *43*(2), 377–386.
- Stevens, K. J., McCabe, C. J., & Brazier, J. E. (2007). Multi-attribute utility function or statistical inference models: A comparison of health state valuation models using the HUI2 health state classification system. *Journal of Health Economics*, *26*(5), 992–1002.
- Stewart, N., Reimers, S., & Harris, A. J. L. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*(3), 687–705.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.
- Trautmann, S. T., & van de Kuilen, G. (2015). Ambiguity attitudes. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (Ch. 3), 89–116. Oxford, UK: Blackwell.
- Toubia, O., Johnson, E., Evgeniou, T., & Delquié, P. (2013). Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science*, *59*(3), 613–640.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, *59*(S4), S251–S278.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.
- Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review*, *80*, 204–217.
- Viscusi, W. K., & Magat, W. A. (1992). Bayesian decisions with ambiguous belief aversion. *Journal of Risk and Uncertainty*, *5*(4), 371–387.
- von Békésy, G. (1947). A new audiometer. *Acta Otolaryngology*, *35*(5-6), 411–422.
- Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge: Cambridge University Press.
- Wakker, P. P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*(8), 1131–1150.
- Weber, R. A., Camerer, C. F., & Knez, M. (2004). Timing and virtual observability in ultimatum bargaining and “weak link” coordination games. *Experimental Economics*, *7*(1), 25–48.
- Williams Jr., C. A. (1966). Attitudes toward speculative risks as an indicator of attitudes toward pure-risk. *Journal of Risk and Insurance*, *33*(4), 577–586.
- Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). A Python package for adaptive design optimization. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-020-01386-4>.
- Zeisberger, S., Vrecko, D., & Langer, T. (2012). Measuring the time stability of prospect theory preferences. *Theory and Decision*, *72*(3), 359–386.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.