A DEFENSE OF PROSPECT THEORY IN BERNHEIM & SPRENGER'S EXPERIMENT (AL_1)

MOHAMMED ABDELLAOUI[a], CHEN LI[b], PETER P. WAKKER[b], & GEORGE WU[c]
a: HEC Paris and CNRS, Jouy-en-Josas, France, abdellaoui@hec.fr
b: Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands, c.li@ese.eur.nl; Wakker@ese.eur.nl
c: Booth School of Business, University of Chicago, Chicago, USA, wu@chicagobooth.edu
20 August, 2020

Bernheim and Sprenger (2020, *Econometrica*) presented experimental evidence aimed to falsify rank dependence (and, thus, prospect theory). We argue that their experiment captured heuristics and not preferences. The same tests, but with procedures that avoid heuristics, have been done before, and they confirm rank dependence. Many other violations of rank dependence have been published before. Bernheim and Sprenger recommend rank-independent probability weighting with complexity aversion, but this is theoretically unsound and empirically invalid. In view of its many positive results, prospect theory with rank dependence remains the best model of probability weighting and the existing model that works best for applied economics.

JEL-CLASSIFICATION: D81, C91
KEYWORDS: prospect theory; rank dependence; complexity aversion

1. INTRODUCTION

Bernheim & Sprenger (2020) (BS henceforth) claim to have experimentally falsified rank-dependent probability weighting and, hence, Tversky & Kahneman's (1992) (cumulative) prospect theory (CPT). We dispute this claim. The main problem in BS's experiment is that their stimuli are too complex while the stakes are too low. Many preceding papers have argued that such stimuli lead to responses based on heuristics rather than preferences; see §§4-5. Replications are desirable that circumvent these issues with filler questions, visual aids, and larger stakes (outcome *differences*). Fortunately, such replications have already been provided in the literature: Weber & Kirsner (1997) for BS's first experiment (our Eq. 5), and Diecidue, Wakker, & Zeelenberg (2007) for BS's second experiment (our Eq. 7). Both studies confirmed rank dependence, showing that the findings of BS are not robust.[1]

BS suggest a theory of rank-independent weighting. However, such weighting is unsound, more than commonly thought (§2.4). In their experimental measurement, BS overlook that a common power of probability weighting and utility cannot be identified from their stimuli. BS further suggest that the preference functional may depend on the number of outcomes of a lottery. Many papers have discussed this idea (§6.3). Tversky & Kahneman (1992 p. 317), for instance, argued that such dependence, as well as similar effects, do not lend themselves to formal analysis. Indeed, the idea never became popular in economics. BS argue for an aversion to many gains, but most empirical studies find the opposite: a preference for many gains (§6.3).

BS criticize a commonly-used statistical technique, but we dispute their criticism (§6.1). Possibly based in part on that criticism, they do not cite (or cite but do not discuss) much preceding literature. We understand that the rank-dependent stream is too big to completely survey, and thus add several references to preceding violations of rank dependence (§6.2, §6.4). Some are like those of BS but more serious.

Beyond our critique of BS's experiments, we argue that models such as prospect theory should be evaluated by considering the body of tests of these theories, with no single test invalidating a model. Besides the many aforementioned violations, many

[1] It is understandable that papers in field journals of over a decade ago have not been widely known.

more studies have supported rank-dependent probability weighting. Its imperfections notwithstanding, prospect theory is the best model with probability weighting for use in modeling economic phenomena such as insurance or asset pricing (Barberis 2013; Fehr-Duda & Epper 2012). Quiggin (1982 Eq. 10) showed that, under mild assumptions, rank dependence is the only probability weighting model that does not violate stochastic dominance.

In sum, in our critique of BS, we put forth these three considerations – whether the experiment measures preferences or heuristics; how the new empirical evidence adds to prior empirical evidence; and the performance of the model in applied settings – as essential for making sense of empirical evidence and the models that they test.

2. THREE PROBLEMS FOR BS'S TREATMENT OF 1979 PROSPECT THEORY

By

$$(p{:}\,X, q{:}\,Y, 1-p-q{:}\,Z), \tag{1}$$

called *lottery*, we denote a probability distribution on $\mathbb{R}^+$ that assigns probability p to X, probability q to Y, and probability $1-p-q$ to Z ($p \geq 0, q \geq 0, p+q \leq 1$). In what follows, we use BS's notation and terminology as much as possible.[2] BS only consider lotteries with three or fewer outcomes, which are all gains (≥ 0). Fewer outcomes result if some of the probabilities in Eq. 1 are 0. By $u{:}\, \mathbb{R}^+ \to \mathbb{R}^+$ we denote a *utility function* (or value function). It is assumed to be strictly increasing and continuous, with $u(0) = 0$. By $\pi{:}\, [0,1] \to [0,1]$ we denote a *weighting function*. It is assumed to be strictly increasing with $\pi(0) = 0$ and $\pi(1) = 1$. For original prospect theory (PT; Kahneman & Tversky 1979), BS propose the following evaluation of the lottery in Eq. 1:

$$\pi(p)u(X) + \pi(q)u(Y) + \pi(1-p-q)u(Z). \tag{2}$$

---

[2] We do not use BS's notation of lotteries because their use of braces to denote arrays rather than sets is unconventional.

This is an understandable proposal that has often been made in the literature but, and this is the first problem, it is not correct strictly speaking. Eq. 2 has been known as *separable prospect theory*. The correct formula for PT, where we may assume $X > Y > Z$ and $1 - p - q > 0$, is:

$$\pi(p)(u(X) - u(Z)) + \pi(q)(u(Y) - u(Z)) + u(Z). \quad (3)$$

That is, the lowest outcome Z, called riskless by Kahneman and Tversky, should not be weighted. The Appendix explains the background of this formula.

The second problem of BT's rank-independent weighting is that both formulas (Eqs. 2 and 3) are not theoretically sound. In particular, as has often been pointed out, they violate stochastic dominance. We add here that they can lead to violations that are absurd in magnitude. Consider the lottery yielding outcome 0 with probability 0.01, and outcome $1 + j \times 10^{-5}$ with probability 0.01 for $j = 1, \ldots, 99$. The certainty equivalent of the lottery, with the parametric estimates of Tversky & Kahneman (1992) and under the extensions of both Eq. 2 and Eq. 3, is 6.90, which exceeds the maximal outcome of the lottery by a factor of more than 6. This does not make any sense.[3] Further, Rieger & Wang (2008) showed that any extension of the original PT to continuous distributions is problematic, depending on π only through $\pi'(0)$ and depending much on the particular discrete approximations chosen.

One of the main empirical findings concerns the overweighting of extreme outcomes (Fehr-Duda & Epper 2012; l'Haridon & Vieider 2019; Luce & Suppes 1965 §4.3; Starmer 2000). It fits well with rank dependence, but cannot be accommodated with rank-independent weighting. For all the aforementioned reasons, the rank-independent versions of PT have been generally abandoned in favor of rank-dependence (Barberis 2013 p. 174).

BS aimed to measure probability weighting and utility. To do so, they only considered lotteries with one nonzero outcome in their first and second experiment, probably because they used these measurements both in their rank-dependent and rank-independent weighting analyses. However, this gives rise to the third problem: a joint power of probability weighting and utility cannot be identified from these stimuli. Thus, $\pi(p)u(x)$ is empirically indistinguishable from $\pi(p)^r u(x)^r$ for any

[3] Normalizing decision weights, as in BS (p. 1402 top) does not help (Wakker 2010 pp. 275-276).

$r > 0$ (Cohen & Jaffray 1988 Eq. 7a). For this reason, Fehr-Duda & Epper (2012 p. 583) strongly advised against using only such stimuli. BS find that

$$(p{:}\,x, 1-p{:}\,0) \to \frac{p^{0.715}}{(p^{0.715}+(1-p)^{0.715})^{1/0.715}} x^{0.941}$$

fits their data best in Experiment 1. But $\left(\frac{p^{0.715}}{(p^{0.715}+(1-p)^{0.715})^{1/0.715}}\right)^{\frac{1}{0.941}} x$ fits their data equally well. In particular, the power family that they assume for utility can never rule out linear (or convex or concave) utility as best fitting. By taking r above small or large enough, π can be as high or low as desired, violating the inverse-S shape of the probability weighting function that BS claim as optimal and that is commonly found. Similarly, in BS's second experiment, $\left(\frac{p^{0.766}}{(p^{0.766}+(1-p)^{0.766})^{1/0.766}}\right)^{\frac{1}{0.982}} x$ fits the data equally well as their claimed optimal fit. Hence, BS could not really measure their model empirically.

Because of the three aforementioned problems, we will not discuss BS's analyses of PT further, including their rank-independent probability weighting. We instead focus on BS's analyses of rank dependence henceforth.

3. DETERMINISTIC ANALYSIS OF BS'S EXPERIMENTS

This section gives some preparatory mathematical definitions. It also displays an assumption of linear utility that will be central in later discussions and is by itself reasonable. We assume CPT, with the following evaluation of lotteries:

$$(p{:}\,X, q{:}\,Y, 1-p-q{:}\,Z) \to w_X u(X) + w_Y u(Y) + w_Z u(Z). \tag{4}$$

Here, u is as above, and w_X, w_Y, and w_Z are *decision weights*. Decision weights are *rank-dependent*. For example, w_X depends on whether X is the best, middle, or worst outcome. We follow BS in using the term rank informally and in not expressing rank dependence in notation.[4]

BS's first and main experiment concerns indifferences of the form

---

[4] Wakker (2010) formalized ranks and rank dependence.

$$(p{:}X, q{:}Y, 1-p-q{:}Z) \sim (p{:}X, q{:}Y+m, 1-p-q{:}Z-k). \tag{5}$$

X, the common outcome, is varied across BS's main experiment, with m and k so small that the ranking of outcomes does not change (i.e., "comonotonicity" is satisfied). Throughout, $Y = 24$, $Z = 18$, $m = 5$, and $q = 0.3$. Price lists are used to elicit k (called an *equalizing reduction* by BS) for three different values of p: $p = 0.1$, $p = 0.4$, or $p = 0.6$.

BS used seven values of X. In some instances, X is the best ranked outcome ($X = 34$, $X = 32$, or $X = 30$); in other cases, X, which we denote X', is ranked in the middle ($X' = 23$, $X' = 21$, or $X' = 19$). When X is ranked best, the weights for outcomes Y and $Y+m$ are denoted w_Y. When X is ranked middle, they are denoted $w_Y{}'$.

Under linear utility, $\frac{w_Y}{w_Z} = \frac{k}{5}$ and $\frac{w_Y{}'}{w_Z} = \frac{k'}{5}$. The ratio

$$\frac{k'}{k} = \frac{w_Y'}{w_Y} \tag{6}$$

captures the proportional change of the decision weight and, hence, rank dependence. This ratio, or its log, is used in BS's analysis.

BS repeatedly claim that they can have Eq. 6 for all differentiable utility functions. However, this claim is based on marginal rates of substitution involving infinitesimal changes m and k, which cannot be implemented empirically.[5] Empirically, we have to work with the following, reasonable, assumption:

ASSUMPTION 1 [linear utility for moderate outcome changes]. For outcome changes within a small interval $[A, B]$, utility is approximately linear. □

More precisely, for Eq. 6, it can be seen that linearity of utility is used on the interval $[\min\{18-k, 18-k'\}, 18]$. Assumption 1 provides a good approximation for all

---

[5] BS's Footnote 13 even claims validity for infinitesimals for every strictly increasing continuous utility, dispensing with differentiability. However, this is not correct. For singular Cantor-type functions the *positive* right derivatives claimed by BS may exist *nowhere*, let be at the points where needed. See Paradís, Viader, & Bibiloni (2001; their Theorem 3.1 and its proof are also valid for right and left derivatives).

common nonlinear utility functions. Empirical evidence supporting it is also abundant.[6]

BS's second experiment concerned indifferences of the form

$$(p{:}X, q{:}Y, 1-p-q{:}Z) \sim (p{:}X+m, q{:}Y-k, 1-p-q{:}Z-k). \qquad (7)$$

In this experiment, $p = 0.4$ and $q = 0.3$, or $p = 0.6$ and $q = 0.2$, with $Y = 36$, $Z = 18$, and $m = 4$ throughout. Finally, $X = 2, 3, 4, 20, 21, 22, 38, 39$, or 40, with price lists again used to elicit k, which BS again call an *equalizing reduction*. Under linear utility and Eq. 4,

$$w_X = \frac{k}{m+k}\,. \qquad (8)$$

Consider $X = 4$ (with k) and $X' = 20$ (with k'). In the lottery with $X = 4$, X has the worst rank, whereas $X' = 20$ has the middle rank in the corresponding lottery. The rank of Z changed from middle to worst in these two lotteries. Such rank changes affect the decision weight in Eq. 8. BS again capture the change by the ratio $\frac{k'}{k}$.[7] Eq. 8 again uses Assumption 1. More precisely, it uses linearity of utility on the intervals $[\min\{18-k, 18-k'\}, 18]$ and $[\min\{36-k, 36-k'\}, 36]$. As argued before and supported by numerical analyses by BS, this is a reasonable assumption.

4. SMALL PAYOFF CHANGES: RAMSEY'S TRIFLE PROBLEM AND A STATISTICAL PROBLEM

Ramsey (1931) pointed out a difficulty that applies to BS's implementation of Assumption 1, which we call *Ramsey's trifle problem*:

---

[6] See Birnbaum (2008 p. 469), Epper, Fehr-Duda, & Bruhin (2011), Homonoff (2018 p. 182), Kahneman & Lovallo (1993), Lopes & Oden (1999 footnote 1), Luce (2000 p. 86), Marshall (1890 Book III), Pigou (1920 p. 785), Rabin (2000), Savage (1971 p. 786).

[7] More precisely, the ratio of decision weights is $\frac{k'}{k} \times \frac{m+k}{m+k'}$ which is, roughly, a monotonic nonlinear transformation of k'/k. Importantly, it does not affect being larger or smaller than 1. We conjecture that BS still used k'/k, or its log, as index in their analysis for this reason.

> Since it is universally agreed that money has a diminishing marginal utility, if money bets [to measure decision weights (subjective probabilities) through ratios] are to be used, it is evident that they should be for as small stakes as possible. *But then again the measurement is spoiled by introducing the new factor of reluctance to bother about trifles*. [Italics added] [p. 176]

This trifle problem was also pointed out by Samuelson (1959 Footnote 5). In order to approximate infinitesimal changes with perfect linearity, BS took payoff changes m, k that are very small. But these changes became too small to motivate subjects.

In BS's main (first) experiment, subjects completed 28 price lists, with 21 of those (seven values of X and three different sets of probabilities) constituting the elicitation of equalizing reductions, k or k'. Altogether, subjects answered 980 $(21 \times 38 + 7 \times 26)$ questions, most of which involved nearly-identical lotteries. Although subjects earned $26.87 on average, the effective incentive was the possibility of getting $4 extra with probability 0.3 at the cost of a chance at losing a few dollars. Similar numbers and stakes appeared in BS's second experiment. It is inconceivable that subjects, even if only subconsciously in an as-if sense, would do anything near determining preference values of these complex lotteries, several hundreds of times in a row, for such small stakes. Instead, it is likely that subjects, when facing the figures of the stimuli (BS Online Appendix) will quickly recognize the structure of Eq. 5 or 7, develop a simple algebraic heuristic and use that repeatedly (combined with the usual noise) to quickly get through the experiment (§5). Smith (1982) posited a dominance requirement for experimental economics: the rewards should dominate subjective costs. Wilcox (1993) confirmed empirically that good incentives are necessary for complex stimuli. These requirements are violated by BS. von Winterfeldt & Edwards (1982) reviewed a number of studies that showed that subjects use simple strategies in situations with inadequate incentives.

We also note a statistical problem concerning preferences (as opposed to BS's data). BS's §2.1 reports a deterministic CPT preference analysis of their stimuli that uses ratios $\frac{k}{k'}$, where k and k' are small in an absolute sense relative to the other numbers in the stimuli. We note here that $18 - k$ and $18 - k'$ were the values actually elicited. Small relative errors in these give large relative errors in k, k'. Hence, ratios $\frac{k}{k'}$ are very vulnerable to noise. As BS emphasize throughout, it is important to reckon with noise beyond a deterministic analysis. It would have been of special interest to analyze the role of plausible noise *in preferences* in their §2.1. Adding an error term to their CPT values (as in BS's Eq. 5 and Footnote 60) affects the certainty equivalents of the overall lotteries by some dollars. Given the complexity of their three-outcome lotteries and Ramsey's trifle problem, such errors in preference values are plausible. This leads to errors in the measured k, k' that may readily make them approximate 0 (no negative answers were possible). If such errors occur with probabilities exceeding 0.05, then the confidence intervals of the ratios $\frac{k}{k'}$ span the whole $\mathbb{R}^+$. Then BS's analysis may lack the statistical power to reject any hypothesis *about preference*, be it rank dependence or rank independence.

The above statistical analysis was back-of-the-envelope, for illustrative purposes. It shows that a fully elaborated power analysis, based on adding plausible error models *for preferences* to the calibrated CPT models used throughout BS's paper, would have provided useful insights. It would have shown if the claim in their abstract "Conventional calibrations of CPT preferences imply that the percentage change in probability weights should be an order of magnitude larger than we observe," and claimed nonoverlapping confidence intervals (BS p. 1366 middle; p. 1382 l. 12; p. 1388), can hold statistically for noisy-preference calibration models. It would also show whether the variances found in their data may at all represent preferences rather than heuristics, as we argue. In general, power analyses are best done prior to observing data. For brevity, we do not elaborate on them.

5. HEURISTICS IN BS'S EXPERIMENTS

The data that BS found did not exhibit the volatility suggested at the end of the preceding section. BS obtained stable patterns giving statistical power and tight confidence intervals. Our claim is that this is because their experiment did not measure preferences, and did not, even in an as-if sense, speak to Eqs. 2, 3, or 4. Instead, subjects faced with hundreds of choices of complex and nearly-identical lotteries, for a one-time trifle reward received with some probability, develop simple algebraic heuristics to get through the experiment.

Many studies have shown that multiple repetitions of complex tasks can lead to stable but invalid patterns, in our case heuristics instead of preferences. Ariely, Loewenstein, & Prelec (2001) call it coherent arbitrariness, while Loomes, Starmer, & Sugden (2003) call it the shaping hypothesis. See also Baron et al. (2001 p. 3 l. -2), Carlin (1992 p. 219), Dolan & Stalmeier (2003), and Hardisty et al. (2013).

This main heuristic in BS's first and main experiment, based on Eq. 5, was cancellation: subjects ignored the common outcome X, not because of preference, but only as a heuristic to simplify the experimental task. It precludes rank dependence. BS (p. 1366) acknowledge this problem, citing Wu (1994) and Weber & Kirsner (1997). Cancellation has been widely documented in the literature (see below). Weber & Kirsner (1997 top of p. 57) showed that Wakker, Erev, & Weber (1994) suffered from cancellation, explaining the absence of rank dependence there. All these papers used stimuli as in BS's first experiment (Eq. 5). In a treatment using such stimuli while avoiding cancellation (by asking for pricing rather than direct choice) and other heuristics, Weber & Kirsner did find rank dependence. The problem of cancellation in Wakker, Erev, & Weber's design reappears in BS's first experiment.

A general finding is that, the more saliently the common outcome is displayed, the stronger cancellation is.[8] In the stimuli of BS's first experiment, throughout and invariably, the most left column of lotteries displays the common outcome, each time spanning the whole page (BS Online Appendix). This is as salient as it can be. Heuristics were further facilitated by the following features of the stimuli, that can be inferred from the experimental instructions provided online. Invariably, the middle column displayed the outcomes 24 (for the left lottery) and 29 (for the right lottery), always with the same probability 0.3. Further, the right column always displayed outcomes \$18 (for the left lottery) and $\$18 - k$ (for the right lottery), with the same probability vector. Probabilities of those outcome always decreased in the same order in all blocks of three.

To avoid cancellation, BS (§5.3) carried out a second, smaller experiment, based on Eq. 7. Now there is no common outcome to be cancelled. They also had a bigger variation in outcomes X, which has two advantages. First, it makes it harder for subjects to develop some simple heuristics that ignore nonvarying outcomes. Second,

[8] We give one reference from every decade: Kahneman & Tversky (1979 p. 274—on their Figures 1 versus 2); Keller (1985); Kashima & Maher (1995); Birnbaum (2008 p.481 ff.) ; Schneider, Leland, & Wilcox (2018); Blavatskyy, Ortmann, & Panchenko (2020 —compound vs. lottery).

middle ranks are not only changed into best ranks, where rank dependence is known to be weak (§6.2), but also into worst ranks, where rank dependence is known to be stronger (Fehr-Duda & Epper 2012 end of §2; Wakker 2010 p. 227; Weber & Kirsner 1997 p. 58).

The format in BS's second experiment has not been used much before and, hence, less is known about presence or absence of heuristics. Despite the aforementioned advantages, we still think that heuristics were measured and not preferences, because many problems of BS's Experiment 1 remain in their Experiment 2. The complexity of the lotteries has worsened due to the absence of a common outcome. This augments Ramsey's trifle problem. The statistical problem of ratios of small numbers also remains. The layout of the stimuli (their Online Appendix, Figure 5) facilitates the heuristics with, again, the same format for hundreds of choices over several pages, again inducing coherent arbitrariness.

It is, for instance, heroic to think that, for the many complex lotteries and trifle rewards, subjects would incorporate the separate values $Y = 24, Y = 24 - k,\ Z = 18,\ Z = 18 - k,\ q$, and $1 - p - q$ from Eq. 7 into their valuation, even in an as-if sense. Instead, if the values were subsumed into one concept, "lose k otherwise,"[9] there is no perception of the ranking of outcomes and no scope for rank dependence.

Diecidue, Wakker, & Zeelenberg (2007; DWZ henceforth) used the same equalizing reductions as in BS's second experiment, i.e., indifferences as in our Eq. 7; see their Eq. 3.2. They thus avoided the common outcomes in Wakker, Erev, & Weber (1994) that had been criticized by Weber & Kirsner (1997). They also used Assumption 1 to obtain nonparametric estimations of decision weights, and their primary purpose was also to test rank dependence (main hypothesis on p. 185; p. 195 2nd para). Their amounts m, k were considerably larger than in BS (m never below Dfl. 20, which in present value would be about \$20). We discuss this difference. The exact validity for general utility in BS's theorems only concerns infinitesimal m, k. In experiments, as in BS and DWZ, one needs to take small discrete m, k, small enough to have good approximations. However, if m, k are too small, as in BS, then one runs into Ramsey's trifle problem, and the statistical problem at the end of §4. Hence, m, k

[9] Given probability p for X and $X + 4$, loss k automatically occurs with complementary probability $1 - p$.

have to be larger (but not too much), as in DWZ. And one needs Assumption 1, which is still reasonable for the amounts considered (referenced above). This is the only way in which BS's approach with equalizing reductions can be used, and this is what DWZ did.

In the beginning of §3, DWZ explain that three-outcome prospects are too difficult to evaluate in general; see also DWZ (p. 181, 3rd & 4th para). Hence, they used a visual design (their Figure 1) to facilitate these choices,[10] developed in extensive pilot studies with debriefings to identify and then avoid the major heuristics used by subjects (see their p. 188 last two paras; p. 194 last para; p. 195 last para). They used filler questions, and more variations in outcomes, to further reduce heuristics. Following Weber & Kirsner (1997), DWZ minimized all aforementioned problems in the design of BS.[11] Podsakoff, MacKenzie, & Podsakoff (2012) survey general techniques for reducing measurement-instrument biases. DWZ emphasized that avoiding heuristics is desirable to increase statistical power (p. 188 last line; p. 195 3rd para). Their findings supported rank dependence. That is, there were significant changes of decision weight if ranks of events changed (p. 192 last para), in agreement with other findings of rank dependence in the literature.

DWZ also found violations of comonotonic independence, that is, of rank dependence. Decision weights sometimes changed even though ranks did not. BS's finding is reversed in the sense that their decision weights did not change even if ranks did. Whereas BS's finding can be taken as a special case of CPT (expected utility), DWZ's finding is more serious because it cannot be reconciled with any version of CPT. Wu (1994) showed large violations of rank-dependence by testing

---

[10] This is commonly done for lotteries with three or more outcomes, by Weber & Kirsner (1997) and most others. Lola Lopes, specialized in multi-outcome lotteries, developed special visual designs (Lopes & Oden 1999; Fennema & Wakker 1997).

[11] Two further differences between DWZ and BS are as follows. First, DWZ used events with unknown probabilities rather than probabilities. This does not affect the theoretical working of rank dependence. See DWZ (p. 185 ll. 7-8) and Wakker (2010 Figure 7.4.1 versus 10.4.1). Second, an improvement of one prospect was not compensated by worsening that same prospect elsewhere, but instead by improving the other prospect. This avoids confounds due to differences between improvements and worsenings.

comonotonic independence. In particular, he showed that heuristics such as cancellation can sometimes generate patterns inconsistent with prospect theory.

It could be argued that the different findings of DWZ regarding rank dependence are because they considered uncertainty rather than risk. That deviations from expected utility are rank dependent for uncertainty but not for risk. However, we find this explanation implausible. We think that the differences are because DWZ's design, unlike BS's design, reduced heuristics and errors so that they achieved statistical power to test *preference* conditions.

6. FURTHER COMMENTS

6.1. BS'S CRITICISM OF COMMON STATISTICAL TESTS

BS (§2.3) criticize tests that compare numbers of violations of predictions (preference axioms), which are commonly used throughout decision theory and in other empirical sciences. BS erroneously claim that such tests would not assume any ("parametric") noise model, writing:

> These types of frequency comparisons raise two difficulties, both stemming from the fact that the results are difficult to interpret without a parametric model of noisy choice. First, the premise of the approach—that violation frequencies are *necessarily* higher for invalid axioms—is flawed. [italics added] [p. 1376]

See also BS (p. 1367 l. -3 and p. 1376 2nd para). However, the common tests are statistical, and statistical tests always assume a noise model, contrary to BS's "necessarily" claim. These tests only assume that the aforementioned higher frequencies are likely, not necessary. Following up, BS claim the following two difficulties.

The first claimed difficulty concerns a counterexample where all choices testing one preference axiom are close to indifference and subject to much noise, whereas those for the other preference axiom are all far from indifference, with little noise. BS claim that this would invalidate the common tests. We argue that one cannot disregard a whole stream of literature based on one artificial counter-example. For every data set and statistical analysis based on a noise model, one can specify an alternative noise model that invalidates the analysis. However, to serve as a good counterexample, the alternative noise model should be plausible. Every good test based on counting violations was done in a design where BS's first difficulty was not plausible. Note also that there have been many such tests in the literature. Even if one by accident was as in BS's example, then this does not invalidate the whole literature based on it. We finally point out that, even under common designs with plausible error theories, unlikely data may arise as in BS's example, or other kinds of unlikely data, leading to errors of type I or II. Statistical tests never claim that such errors are "necessarily" absent; only, that they are unlikely. P-values, powers, Bayes factors, and so on capture such unlikely exceptions.

BS's second claimed difficulty concerns the example at the end of their Online Appendix B. In this example, they assume rank dependence that "inadvertently" is too weak to affect optimal choice between any of the stimuli considered, and trembling-hand errors depending only on preference and not on utility differences. Then CPT and expected utility have identical predictions in this design. Tests based on numbers of violations then indeed have no power to distinguish. But then, neither does any other test. This trivial example cannot serve as a criticism of any test. We were unable to understand BS's description of this example in their main text (p. 1376): "For any given degree of rank dependence, one can construct simple examples (with constant "distance to indifference") in which the differential between violation frequencies falls anywhere between zero and unity." BS argue that they have refuted some widely used statistical tests and a whole stream of literature based on it. However, their criticisms are unfounded.

6.2. Weakness of rank dependence in Longshot Effect

In their main experiment, BS consider changes in decision weights only when the rank changes from middle to best. It is well-known that rank dependence is not strong there (DWZ p. 185 ll. 4-6; DWZ p. 197 l. 7). The prevailing finding is that the weights then increase, consistent with inverse-S probability weighting and the longshot effect. Stronger rank dependence occurs when ranks change from middle to worst, consistent with the certainty effect. As for the change in rank considered by BS, quite some studies found that the increase of decision weight is weak or absent. Even, several studies found the opposite effect to be prevailing, of decreasing decision weight, consistent with pessimistic probability weighting. See van de Kuilen & Wakker (2011). Their Footnotes 7 & 8 survey the many other papers that found this opposite effect. In view of this literature, finding no effect of rank dependence in BS's first experiment is no surprise anyhow.

6.3. Complexity Seeking Instead of Aversion

BS favor adding a component to risk theory that reckons with the number of outcomes of a lottery, to capture complexity aversion. Many authors have investigated proposals of this sort. Neilson (1992) proposed a formal model, but it was tested unsuccessfully by Humphrey (2001). Related models received some attention in psychology (Birnbaum 2008 p. 481; Krantz et al. 1971 Ch. 8; Luce 2000). Tversky & Kahneman (1992 p. 317) were pessimistic about modeling this phenomenon:

> Despite its greater generality, the cumulative functional is unlikely to be accurate in detail. We suspect that decision weights may be sensitive to the formulation of the prospects, as well as to the *number*, the spacing and the level of outcomes. … The present theory can be generalized to accommodate such effects but it is questionable whether the gain in descriptive validity, achieved by giving up the separability of values and weights, would justify the loss of predictive power and the cost of increased complexity. … The heuristics of choice do not readily lend themselves to formal analysis because their application depends on the formulation of the problem, the method of elicitation, and the context of choice. [italics added]

We agree with this pessimism, and this old model never caught on in economics to our best knowledge. The volatility of empirical findings adds to our pessimism. In a literature review (Online Appendix), we found two additional empirical studies confirming complexity aversion, but seven studies finding the opposite, complexity seeking. Thus, the prevailing empirical finding is opposite to BS's model.

6.4. Further Preceding Falsifications of Prospect Theory

Given that prospect theory is the most tested risk theory, besides being most confirmed, it is also most violated. The keyword "PT falsified" in Wakker (2020) gives 49 papers.

We call special attention to a finding of Birnbaum & McIntosh (1996), which is not cited by BS. It was confirmed in several follow-up studies by Birnbaum and colleagues, surveyed by Birnbaum (2008 pp. 484-487), and found independently by Humphrey & Verschoor (2004). This finding concerns lotteries of the same format as in Eq. 5, i.e., as in BS's first experiment, with the common outcome X moved to test rank dependence. Prospect theory predicts that weights increase if ranks change from middle to best or worst. BS quantitatively find no change in decision weighs. The aforementioned studies attempted to avoid heuristics and found in fact a stronger deviation: a decrease in weight.

The strongest counterexample to rank dependence that we are aware of is Machina's (2009) reflection example, confirmed empirically by l'Haridon & Placido (2010). Ending on a positive note, Fehr-Duda & Epper (2012) and Barberis (2013) provide surveys with many useful applications of rank dependence. See also the impressive data sets of l'Haridon & Vieider (2019) and Ruggeri et al. (2020), and DellaVigna's (2018) review in the context of structural behavioral economics. A psychological justification for rank-dependent decision weights is based on cognitive attention: individuals tend to attend to extreme outcomes (Weber & Kirsner 1997; Pachur et al. 2018).

7. CONCLUSION

We have offered a critique of BS's experiments and analyzes. While we applaud BS's call for more investigations of rank dependence, we attribute their observed stability of equalizing reductions to subjects' use of heuristics, not a failure of rank dependence as claimed by BS. Every study should be interpreted relative to a large set of admittedly mixed empirical results. Prospect theory is an imperfect theory, as will be every theory that aims to make sense of the complex problem of how people make decisions. Nevertheless, for now, we see prospect theory as the most tractable and best performing model with nonlinear probability weighting.

APPENDIX A: BACKGROUND OF PROSPECT THEORY FORMULA OF 1979

Kahneman & Tversky (1979) only wrote Eq. 3 explicitly for at most two nonzero outcomes, i.e., when either $Z = 0$ (their Eq. 1) or $p = 0$ (their Eq. 2). This has led to much confusion about Eqs. 2 and 3 in the literature. We quote their verbal description to show that Eq. 3 is correct.

> ... prospects are segregated into two components: (i) the riskless component, i.e., the minimum gain ... which is certain to be obtained or paid; (ii) the risky component, i.e., the additional gain[s] ... which is[are] actually at stake.... That is, the value of a strictly positive ... prospect equals the value of the riskless component plus the value-difference between the outcomes, multiplied by the weight associated with the more extreme outcome[s]. The essential feature ... is that a decision weight is applied to the value difference ... which represents the risky component of the prospect, but not to … the riskless component. (Kahneman & Tversky. 1979 p. 276).

We added the texts between square brackets to extend to plurality, with multiple non-minimal outcomes.

From the case $p = 0$ it appears that Eq. 2 above, claimed by BS, cannot be correct, as the riskless outcome Z should not be weighted. Eq. 3 is the only natural extension of the formulas given by Kahneman & Tversky. That Eq. 3 is correct also appears from Kahneman & Tversky (1975 p. 18), a first working paper version of their 1979 paper. They do explicitly write the formula of PT for multiple outcomes there. They treat the value function somewhat differently than in their 1979 paper, taking utilities of differences rather than differences of utilities.[12] But they treat probability weighting as in Eq. 3. They also emphasize (their p. 12) that the riskless outcome should be as in our Eq. 3 (not weighted) rather than in our Eq. 2 (where it is weighted). Whenever relevant, they pointed out that the riskless outcome should not be weighted (Tversky & Kahneman 1981 Footnote 5; Tversky & Kahneman 1986 Footnote 2).

BS (footnote 11) cite Camerer & Ho (1994) for Eq. 2. However, Camerer & Ho (1994) used a different term, separable prospect theory, for Eq. 2. Their endnote 16 pointed out that it deviates from prospect theory for strictly positive lotteries. BS (footnote 11) also cite Fennema & Wakker (1997) for Eq. 2. However, Fennema & Wakker (p. 54) pointed out that this equation should be used for mixed prospects, which assign positive probabilities to both gains and losses. Those are among what Kahneman & Tversky called regular prospects, and then the analog of Eq. 2 is indeed correct (Kahneman & Tversky 1975 p. 18).

---

[12] The latter is preferable because it can be applied to nonquantitative outcomes.

REFERENCES

Ariely, Dan, George F. Loewenstein, & Drazen Prelec (2001) " 'Coherent Arbitrariness': Stable Demand Curves without Stable Preferences," *Quarterly Journal of Economics* 118, 73–106.

Barberis, Nicholas (2013) "Thirty Years of Prospect Theory in Economics: A Review and Assessment," *Journal of Economic Perspectives* 27, 173–195.

Baron, Jonathan, Zhijun Wu, Dallas J. Brennan, Christine Weeks, & Peter A. Ubel (2001) "Analog Scale, Magnitude Estimation, and Person Trade-Off as Measures of Health Utility: Biases and Their Correction," *Journal of Behavioral Decision Making* 14, 17–34.

Bernheim, B. Douglas & Charles Sprenger (2020) "On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting," *Econometrica* 88, 1363–1409.

Birnbaum, Michael H. (2008) "New Paradoxes of Risky Decision Making," *Psychological Review* 115, 463–501.

Birnbaum, Michael H. & William R. McIntosh (1996) "Violations of Branch Independence in Choices between Gambles," *Organizational Behavior and Human Decision Processes* 67, 91–110.

Blavatskyy, Pavlo, Andreas Ortmann, & Valentyn Panchenko (2020) "On the Experimental Robustness of the Allais Paradox." *American Economic Journal: Microeconomics*, forthcoming.

Camerer, Colin F. & Teck-Hua Ho (1994) "Violations of the Betweenness Axiom and Nonlinearity in Probability," *Journal of Risk and Uncertainty* 8, 167–196.

Carlin, Paul S. (1992) "Violations of the Reduction and Independence Axioms in Allais-Type and Common-Ratio Experiments," *Journal of Economic Behavior and Organization* 19, 213–235.

Cohen, Michèle & Jean-Yves Jaffray (1988) "Certainty Effect versus Probability Distortion: An Experimental Analysis of Decision Making under Risk," *Journal of Experimental Psychology: Human Perception and Performance* 14, 554–560.

DellaVigna, Stefano (2018) "Structural Behavioral Economics." *In* B. Douglas Bernheim, Stefano DellaVigna, & David Laibson (eds.) *Handbook of Behavioral Economics; Volume 2*, 613–723, Elsevier, Amsterdam.

Diecidue, Enrico, Peter P. Wakker, & Marcel Zeelenberg (2007) "Eliciting Decision Weights by Adapting de Finetti's Betting-Odds Method to Prospect Theory," *Journal of Risk and Uncertainty* 34, 179–199.

Dolan, Paul & Peep Stalmeier (2003) "The Validity of Time Trade-Off Values in Calculating QALYs: Constant Proportional Time Trade-Off versus the Proportional Heuristic," *Journal of Health Economics* 22, 445–458.

Epper, Thomas, Helga Fehr-Duda, & Adrian Bruhin (2011) "Viewing the Future through a Warped Lens: Why Uncertainty Generates Hyperbolic Discounting," *Journal of Risk and Uncertainty* 43, 163–203.

Fehr-Duda, Helga & Thomas Epper (2012) "Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences," *Annual Review of Economics* 4, 567–593.

Fennema, Hein & Peter P. Wakker (1997) "Original and Cumulative Prospect Theory: A Discussion of Empirical Differences," *Journal of Behavioral Decision Making* 10, 53–64.

Hardisty, David J., Katherine F. Thompson, David H. Krantz, & Elke U. Weber (2013) "How to Measure Time Preferences: An Experimental Comparison of Three Methods," *Judgment and Decision Making* 8, 214–235.

Homonoff, Tatiana A. (2018) "Can Small Incentives Have Large Effects? The Impact of Taxes versus Bonuses on Disposable Bag Use," *American Economic Journal: Economic Policy* 10, 177–210.

Humphrey, Steven J. (2001) "Are Event-Splitting Effects Actually Boundary Effects?," *Journal of Risk and Uncertainty* 22, 79–93.

Humphrey, Steven J. & Arjan Verschoor (2004) "Decision-Making under Risk among Small Farmers in East Uganda," *Journal of African Economies* 13, 44–101.

Kahneman, Daniel & Dan Lovallo (1993) "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking," *Management Science* 39, 17–31.

Kahneman, Daniel & Amos Tversky (1975) "Value Theory: An Analysis of Choices under Risk," paper presented at the ISRACON conference on Public Economics, Jerusalem, 1975.

Kahneman, Daniel & Amos Tversky (1979) "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263–291.

Kashima, Yoshihisa & Patrick Maher (1995) "Framing of Decisions under Ambiguity," *Journal of Behavioral Decision Making* 8, 33–49.

Keller, L. Robin (1985) "An Empirical Investigation of Relative Risk Aversion," *IEEE Transactions on systems, Man, and Cybernetics*, SMC-15, 475–482.

Krantz, David H., R. Duncan Luce, Patrick Suppes, & Amos Tversky (1971), "*Foundations of Measurement, Vol. I (Additive and Polynomial Representations).*" Academic Press, New York. (2nd edn. 2007, Dover Publications, New York.)

l'Haridon, Olivier & Laetitia Placido (2010) "Betting on Machina's Reflection Example: An Experiment on Ambiguity," *Theory and Decision* 69, 375–393.

l'Haridon, Olivier & Ferdinand Vieider (2019) "All over the Map: A Worldwide Comparison of Risk Preferences," *Quantitative Economics* 10, 185–215.

Loomes, Graham, Chris Starmer, & Robert Sugden (2003) "Do Anomalies Disappear in Repeated Markets?," *Economic Journal* 113, C153–C166.

Lopes, Lola L. & Gregg C. Oden (1999) "The Role of Aspiration Level in Risky Choice: A Comparison of Cumulative Prospect Theory and SP/A Theory," *Journal of Mathematical Psychology* 43, 286–313.

Luce, R. Duncan (2000) "*Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*." Lawrence Erlbaum Publishers, London.

Luce, R. Duncan & Patrick Suppes (1965) "Preference, Utility, and Subjective Probability." *In* R. Duncan Luce, Robert R. Bush, & Eugene Galanter (eds.) *Handbook of Mathematical Psychology*, Vol. III, 249–410, Wiley, New York.

Machina, Mark J. (2009) "Risk, Ambiguity, and the Rank-Dependence Axioms," *American Economic Review* 99, 385–392.

Marshall, Alfred (1890) "*Principles of Economics.*" 8th edn. 1920 (9th edn. 1961), MacMillan, New York.

Neilson, William S. (1992) "Some Mixed Results on Boundary Effects," *Economics Letters* 39, 275–278.

Pachur, Thorsten, Michael Schulte-Mecklenbeck, Ryan O. Murphy, & Ralph Hertwig (2018) "Prospect Theory Reflects Selective Allocation of Attention," *Journal of Experimental Psychology: General* 147, 147–169.

Paradís, Jaumnine, Pelegri Viader, & Lluis Bibiloni (2001) "The Derivative of Minkowski's ?(x) Function," *Journal of Mathematical Analysis and Applications* 253, 107–125.

Pigou, Arthur C. (1920) "*The Economics of Welfare.*" (edn. 1952: MacMillan, London.)

Podsakoff, Philip M., Scott B. MacKenzie, & Nathan P. Podsakoff (2012) “Sources of Method Bias in Social Science Research and Recommendations on How to Control It,” *Annual Review of Psychology* 63, 539–69.

Quiggin, John (1982) “A Theory of Anticipated Utility,” *Journal of Economic Behaviour and Organization* 3, 323–343.

Rabin, Matthew (2000) “Risk Aversion and Expected-Utility Theory: A Calibration Theorem,” *Econometrica* 68, 1281–1292.

Ramsey, Frank P. (1931) “Truth and Probability.” *In* Richard B. Braithwaite (ed.), *The Foundations of Mathematics and other Logical Essays,* 156–198, Routledge and Kegan Paul, London.

Reprinted in Henry E. Kyburg Jr. & Howard E. Smokler (1964, eds.) *Studies in Subjective Probability*, 61–92, Wiley, New York. (2nd edn. 1980, Krieger, New York.)

Rieger, Marc Oliver & Mei Wang (2008) “Prospect Theory for Continuous Distributions,” *Journal of Risk and Uncertainty* 36, 83–102.

Ruggeri, Kai, Sonia Alí, Mari Louise Berge, et al. (2020) “Replicating Patterns of Prospect Theory for Decision under Risk,” *Nature Human Behavior* 4. 622 –633.

Samuelson, Paul A. (1959) “The St. Petersburg Paradox as a Divergent Double Limit,” *International Economic Review* 1, 31–37.

Schneider, Mark, Jonathan W. Leland, & Nathaniel T. Wilcox (2018) “Ambiguity Framed,” *Journal of Risk and Uncertainty* 57, 133–151.

Tversky, Amos & Daniel Kahneman (1981) “The Framing of Decisions and the Psychology of Choice,” *Science* 211, 453–458.

Tversky, Amos & Daniel Kahneman (1986) “Rational Choice and the Framing of Decisions,” *Journal of Business* 59, S251–S278.

Tversky, Amos & Daniel Kahneman (1992) “Advances in Prospect Theory: Cumulative Representation of Uncertainty,” *Journal of Risk and Uncertainty* 5, 297–323.

Savage, Leonard J. (1971) “Elicitation of Personal Probabilities and Expectations,” *Journal of the American Statistical Association* 66, 783–801.

Smith, Vernon L. (1982) “Microeconomic Systems as an Experimental Science,” *American Economic Review* 72, 923–955.

Starmer, Chris (2000) "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature* 38, 332–382.

van de Kuilen, Gijs & Peter P. Wakker (2011) "The Midweight Method to Measure Attitudes toward Risk and Ambiguity," *Management Science* 57, 582–598.

von Winterfeldt, Detlof & Ward Edwards (1982) "Costs and Payoffs in Perceptual Research," *Psychological Bulletin* 91, 609–622.

Wakker, Peter P. (2010) "*Prospect Theory: For Risk and Ambiguity*." Cambridge University Press, Cambridge, UK.

Wakker, Peter P. (2020), "Annotated Bibliography." http://personal.eur.nl/wakker/refs/webrfrncs.docx

Wakker, Peter P., Ido Erev, & Elke U. Weber (1994) "Comonotonic Independence: The Critical Test between Classical and Rank-Dependent Utility Theories," *Journal of Risk and Uncertainty* 9, 195–230.

Weber, Elke U. & Britt Kirsner (1997) "Reasons for Rank-Dependent Utility Evaluation," *Journal of Risk and Uncertainty* 14, 41–61.

Wilcox, Nathaniel T. (1993) "Lottery Choice: Incentives, Complexity and Decision Time," *Economic Journal* 103, 1397–1417.

Wu, George (1994) "An Empirical Test of Ordinal Independence," *Journal of Risk and Uncertainty* 9, 39–60.

ONLINE APPENDIX OF

A DEFENSE OF PROSPECT THEORY IN BERNHEIM & SPRENGER'S EXPERIMENT: A REVIEW OF COMPLEXITY AVERSION

MOHAMMED ABDELLAOUI[a], CHEN LI[b], PETER P. WAKKER[b], & GEORGE WU[c]
a: HEC Paris and CNRS, Jouy-en-Josas, France, Abdellaoui@hec.fr
b: Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands, c.li@ese.eur.nl; Wakker@ese.eur.nl
c: Booth School of Business, University of Chicago, Chicago, USA, wu@chicagobooth.edu

20 August, 2020

Bernheim & Sprenger (2020), BS henceforth, suggest that people are usually complexity averse. Complexity here refers only to the number of outcomes of lotteries. Aversion to more comprehensive or different forms of complexity of lotteries has been studied by Armantier & Treich (2016), Bruce & Johnson (1996), Kovarik, Levin, & Wang (2016), Mador, Sonsino, & Benzion (2000), and Sonsino, Benzion, & Mador (2002). Following BS, we focus only on empirical studies that have investigated the number of outcomes. BS cite five papers on complexity aversion in their footnote 70. The first four, Iyengar & Kamenica (2010), Iyengar & Lepper (2000), Iyengar,Jiang, & Huberman (2003), and Sonsino & Mandelbaum (2001), considered a different topic, preference against flexibility (number of available choice options to choose one from). The fifth, Stodder (1997), is on confusions of averages versus marginals and complexity of multiple stage lotteries, which, again, are different topics. (It is also only theoretical, with no data.) Hence, these references will not be considered here.

Complexity aversion here means that, other things equal, people prefer lotteries with few outcomes because they are less complex. We review the literature, focusing on gains, the domain considered by BS. We conclude that the prevailing finding is the opposite. That is, people usually prefer lotteries with many outcomes to few outcomes, and, in this sense, they are complexity seeking.

BS find that certainty equivalents of lotteries $(0.4{:}\,30, 0.6{:}\,20)$ exceed those of $(0.4{:}\,30, 0.3{:}\,20 + \varepsilon, 0.3{:}\,20 - \varepsilon)$ considerably, even for small $\varepsilon > 0$, with similar findings for $(0.6{:}\,30, 0.4{:}\,20)$ versus $(0.3{:}\,30 + \varepsilon, 0.3{:}\,30 - \varepsilon, 0.4{:}\,20)$. By any rational theory, the certainty equivalents should, to the contrary, be almost the same for small ε. On the basis of these two observations, BS conclude that people are complexity averse.

Many studies have tested special preferences for numbers of outcomes, usually considering a pure case: certainty equivalents are measured for different framings of identical lotteries, for instance $(0.4{:}\,30, 0.6{:}\,20)$ versus $(0.4{:}\,30, 0.3{:}\,20\ 0.3{:}\,20)$. Although all rational theories of choice require identical certainty equivalents, experiments find systematic violations. Here a pure effect of perceived number of outcomes occurs. Terms used to designate such violations include boundary effects, violations of coalescing (collapsing), and event/outcome splitting effects. The latter term is sometimes (e.g., in works by Humphrey, Starmer, and Sugden) combined with a directional assumption, being complexity seeking. Birnbaum does not add this directional assumption to this term.

Violations of coalescing can be taken as a special case of the attribute splitting effect (Weber, Eisenführ, & von Winterfeldt 1988), or the part-whole bias (Bateman et al. 1997), or the unpacking effect (Tversky & Koehler 1994). Here splitting something up increases the total weight. This underlies Birnbaum's theories. He studied violations of coalescing most extensively. His RAM and TAX models predict that splitting the best outcome of a lottery improves the lottery, but splitting the worst outcome (also if a gain) worsens it. If one normalizes decision weights to always add to 1, as in Birnbaum's models, then Birnbaum's predictions will hold. Then increasing the weight of the worst outcome indeed worsens the value. If one does not normalize the weights, as in separable prospect theory, then increasing the weights of gains (whether best or worst) improves the value. Combining these ideas suggests a strong preference for event splitting if it concerns the best outcome, and less clear effects for the worst outcome, but probably a preference against. Overall, we can then expect more preference for than against event splitting. In other words, the preceding arguments suggest more complexity seeking than aversion. This is indeed what our literature review finds.

Our literature search is based on searching the terms “boundary,” “collaps,” “coalesc,” “complex,” and “split” in Wakker (2020), where we excluded the certainty effect and followed up on cited papers.

- The following three papers report prevailing complexity aversion: Bernheim & Sprenger (2020), Huck & Weizsäcker (1999), Moffatt, Sitzia, & Zizzo (2015).

- The following seven papers report prevailing complexity seeking: Birnbaum (2005), Birnbaum (2007), Humphrey (1995), Humphrey (2000), Humphrey (2001a), Humphrey (2006), Starmer & Sugden (1993).

- The following five papers report about as much aversion as seeking: Birnbaum (2004), Birnbaum, Schmidt, & Schneider (2017), Schmidt & Seidl (2014), Humphrey (2001b), Weber (2007).

We conclude that the findings on complexity aversion are volatile, but the literature has documented more complexity seeking than aversion for gains.

REFERENCES FOR ONLINE APPENDIX

Armantier, Olivier & Nicolas Treich (2016) “The Rich Domain of Risk,” *Management Science* 62, 1954–1969.

Bateman, Ian J., Alistair Munro, Bruce Rhodes, Chris Starmer, & Robert Sugden (1997) “Does Part-Whole Bias Exist? An Experimental Investigation,” *Economic Journal* 107, 322–332.

Bernheim, B. Douglas & Charles Sprenger (2020) “On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting,” *Econometrica* 88, 1363–1409.

Birnbaum, Michael H. (2004) “Causes of Allais Common Consequence Paradoxes: An Experimental Dissection,” *Journal of Mathematical Psychology* 48, 87–106.

Birnbaum, Michael H. (2005) “Three New Tests of Independence That Differentiate Models of Risky Decision Making,” *Management Science* 51, 1346–1358.

Birnbaum, Michael H. (2007) “Tests of Branch Splitting and Branch-Splitting Independence in Allais Paradoxes with Positive and Mixed Consequences,” *Organizational Behavior and Human Decision Processes* 102, 154–173.

Birnbaum, Michael H., Ulrich Schmidt, & Miriam D. Schneider (2017) “Testing Independence Conditions in the Presence of Errors and Splitting Effects,” *Journal of Risk and Uncertainty* 54, 61–85.

Bruce, Alistair C. & Johnnie E. V. Johnson (1996) “Decision-Making under Risk: The Effect of Complexity on Performance,” *Psychological Reports* 79, 67–76.

Huck, Steffen & Georg Weizsäcker (1999) “Risk, Complexity, and Deviations from Expected-Value Maximization: Results of a Lottery Choice Experiment,” *Journal of Economic Psychology* 20, 699–715.

Humphrey, Steven J. (1995) “Regret Aversion or Event-Splitting Effects? More Evidence under Risk and Uncertainty,” *Journal of Risk and Uncertainty* 11, 263–274.

Humphrey, Steven J. (2000) “The Common Consequence Effect: Testing a Unified Explanation of Recent Mixed Evidence,” *Journal of Economic Behavior and Organization* 41, 239–263.

Humphrey, Steven J. (2001a) “Are Event-Splitting Effects Actually Boundary Effects?,” *Journal of Risk and Uncertainty* 22, 79–93.

Humphrey, Steven J. (2001b) “Non-transitive Choice: Event-Splitting Effects or Framing Effects?,” *Economica* 68, 77–96.

Humphrey, Steven J. (2006) “Does Learning Diminish Violations of Independence, Coalescing and Monotonicity?,” *Theory and Decision* 61, 93–128.

Iyengar, Sheena S., & Emir Kamenica (2010) “Choice Proliferation, Simplicity Seeking, and Asset Allocation,” *Journal of Public Economics* 94, 530–539.

Iyengar, Sheena S., & Mark R. Lepper (2000) “When Choice Is Demotivating: Can One Desire Too Much of a Good Thing?” *Journal of Personality and Social Psychology* 79, 995–1006.

Iyengar, Sheena S., Wei Jiang, & Gur Huberman (2003) “How Much Choice Is Too Much: Contributions to 401(k) Retirement Plans,” Pension Research Council Working Paper 2003-10.

Kovarik, Jaromir, Dan Levin & Tao Wang (2016) “Ellsberg Paradox: Ambiguity and Complexity Aversions Compared,” *Journal of Risk and Uncertainty* 52, 47–64.

Mador, Galit, Doron Sonsino, & Uri Benzion (2000) “On Complexity and Lotteries’ Evaluations — Three Experimental Observations,” *Journal of Economic Psychology* 21, 625–637.

Moffatt, Peter G., Stefania Sitzia, & Daniel John Zizzo (2015) “Heterogeneity in Preferences towards Complexity,” *Journal of Risk and Uncertainty* 51, 147–170.

Schmidt, Ulrich & Christian Seidl (2014) “Reconsidering the Common Ratio Effect: The Roles of Compound Independence, Reduction, and Coalescing,” *Theory and Decision* 77, 323–339.

Sonsino, Doron, Uri Benzion, & Galit Mador (2002) “The Complexity Effects on Choice with Uncertainty—Experimental Evidence,” *Economic Journal* 112, 936–965.

Sonsino, Doron, & Marvin Mandelbaum (2001) “On Preference for Flexibility and Complexity Aversion: Experimental Evidence,” *Theory and Decision* 51, 197–216.

Starmer, Chris & Robert Sugden (1993) “Testing for Juxtaposition and Event-Splitting Effects,” *Journal of Risk and Uncertainty* 6, 235–254.

Stodder, James (1997) “Complexity Aversion: Simplification in the Herrnstein and Allais Behaviors,” *Eastern Economic Journal* 23, 1–15.

Tversky, Amos & Derek J. Koehler (1994) “Support Theory: A Nonextensional Representation of Subjective Probability,” *Psychological Review* 101, 547–567.

Wakker, Peter P. (2020), “Annotated Bibliography.” http://personal.eur.nl/wakker/refs/webrfrncs.docx

Weber, Bethany J. (2007) “The Effects of Losses and Event Splitting on the Allais Paradox. Judgment in Decision Making,” *Judgment and Decision Making* 2, 115–125.

Weber, Martin, Franz Eisenführ, & Detlof von Winterfeldt (1988) “The Effects of Splitting Attributes on Weights in Multiattribute Utility Measurement,” *Management Science* 34, 431–445.