A criticism of Bernheim & Sprenger's (2020) tests of rank dependence[1]

Peter P. Wakker
Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands,
Wakker@ese.eur.nl
October, 2022
Journal of Behavioral and Experimental Economics, forthcoming

ABSTRACT

Bernheim and Sprenger (2020, *Econometrica*; SB) claimed to experimentally falsify rank dependence in prospect theory. This comment criticizes SB's results and novelty claims. Their experiments only captured well-known heuristics and not genuine preferences. Many more falsifications of rank dependence have been acknowledged before, where SB's equalizing reductions also have been used before. SB thought to identify probability weighting and utility where they are unidentifiable, which invalidates all SB's related claims. SB used an incorrect formula of original prospect theory. SB's suggested alternative of rank-independent probability weighting with dependence on number of outcomes (their "complexity aversion;" a misnomer) has long been discarded.

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

[1] This paper heavily builds on Abdellaoui, Li, Wakker, & Wu (2020), often verbatim. However, I alone am responsible for errors.

1. Introduction

Bernheim & Sprenger (2020) (SB henceforth[2]) claimed to falsify rank-dependent probability weighting. Rank dependence was introduced independently by Quiggin (1982) and Schmeidler (1989; first version 1982). Tversky & Kahneman (1992) used it to improve their prospect theory (CPT). This comment criticizes SB, as well as the follow-up paper Bernheim, Royer, & Sprenger (2022) (§9). This comment focuses on criticisms relevant to the main conclusions. Online Appendix OA.4 (Wakker 2022c) lists 15 further inaccuracies in SB.

Several papers, not cited by SB, reported violations of rank dependence and prospect theory more serious than those claimed by SB (§5; §8). Wakker's (2022a) keyword "PT falsified" gives 59 violations, reproduced here in Online Appendix OA.5 (Wakker 2022c). Still, there are so many more positive findings that CPT is the most popular descriptive risk theory available today; no better alternative is available.[3]

SB's main experimental problem is that their stimuli are too complex while the stakes (payoff *differences*) are too low. Several preceding papers considered tests as in SB but avoided these problems and then did find rank dependence (§§4-5). Hirshman & Wu (2022) will provide a replication of SB correcting their experimental mistakes. SB claimed to introduce a new measurement method of decision weights. However, they were preceded by Diecidue, Wakker, & Zeelenberg (2007), not cited by SB, who showed that linear utility is needed for this method, contrary to SB's claim of general validity (§5).

SB suggested probability weighting of original prospect theory (Kahneman & Tversky 1979), which they called rank-independent, as an improvement over rank-dependent weighting. Unfortunately, they used an incorrect formula (§2.1). Further, this approach has long been known to be too problematic (§2.2). SB (their §3.2) thought to identify probability weighting and utility from stimuli where they were in fact unidentifiable (§2.3), invalidating all their related claims.

---

[2] We avoid the abbreviation BS for linguistic reasons.

[3] See Barberis (2013 p. 2068, finance); Dean & Ortoleva (2017 p. 386, economics); Fehr-Duda & Epper (2012 §6, economics); Murphy & ten Brincke (2018 p. 309); Pachur et al. (2018 p. 408, psychology); Ruggeri et al. (2020 abstract, general).

SB suggested, as a second improvement, preference functionals that depend on the number of outcomes in a lottery, which they call complexity aversion. However, this idea has long been known not to work (§6). Thus, SB did not provide a viable alternative to the rank-dependent weighting that they criticize.

2. Three problems for SB's treatment of 1979 prospect theory

By

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

called *lottery*, we denote a probability distribution over *outcomes* (monetary gains; $\mathbb{R}^+$) that assigns probability p to outcome X, probability q to outcome Y, and probability $1-p-q$ to outcome Z ($p \geq 0, q \geq 0, p+q \leq 1$).[4] SB only considered lotteries with three or fewer outcomes. 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 strictly increasing and continuous, with $u(0) = 0$. By $\pi{:}\,[0,1] \to [0,1]$ we denote a *weighting function*. It is strictly increasing and continuous with $\pi(0) = 0$ and $\pi(1) = 1$.

2.1. Incorrect formula

SB defined *rank-independent probability weighting* as the following evaluation of the lottery in Eq. 1:

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

Contrary to SB's claims, and as explained by Wakker (2022b), this is not original prospect theory of Kahneman & Tversky 1979).[5] It has been known as *separable*

---

[4] I deviate from SB's notation of lotteries because their use of braces to denote arrays rather than sets violates common conventions. In general, I follow their terminology and notation as much as possible, sometimes reluctantly, so as to facilitate communication.

[5] SB (footnote 11) cited Camerer & Ho (1994) for Eq. 2. However, Camerer & Ho's endnote 16 pointed out that Eq. 2 deviates from prospect theory for strictly positive lotteries. SB (footnote 11) also

prospect theory. When Nilsson, Rieskamp, & Wagenmakers (2020) discovered that an earlier paper by them had erroneously used Eq. 2 instead of original prospect theory, they and the Journal of Mathematical Psychology took a principled stance and published a correction using the right formula.

2.2. Problematic violations of stochastic dominance

Rank-independent weighting (our Eq. 2) was popular in the psychological literature until the 1980s, but was abandoned when Fishburn (1978) discovered that it violates stochastic dominance, as does original prospect theory. However, it has not always been understood that these violations are extreme and absurd, so much that the theory is unacceptable both theoretically and empirically. To illustrate this point, Wakker (2022b) considered the following lottery:

$$(0.01{:}\,1 + 1 \times 10^{-5},\ \ldots,\ 0.01{:}\,1 + 99 \times 10^{-5},\ 0.01{:}\,0).$$

With the parametric estimates of Tversky & Kahneman (1992)[6] and under the natural extension of Eq. 2 and also of original prospect theory to multiple outcomes (Wakker 2022b), the certainty equivalent of this lottery is 6.9, exceeding the maximal outcome of the lottery more than 6-fold. This does not make any sense. The basic problem of rank-independent weighting is that the requirements for lotteries with few outcomes are entirely incompatible with those for many outcomes.[7]

Rieger & Wang (2008) further showed that extensions of rank-independent weighting to continuous distributions are not well possible, confirming preceding conjectures in the literature (Quiggin 1982 p. 330).[8] Some studies suggested that rank-independent weighting may work similarly well as CPT for fitting student-lab choices

---

cited Fennema & Wakker (1997) for Eq. 2. However, Fennema & Wakker (p. 54) pointed out that they use Eq. 2 only for mixed loteries, which assign positive probabilities to both gains and losses. Then Eq. 2 does agree with original prospect theory (Wakker 2022b).

[6] Because original and new prospect theory agree on the two-outcome gain lotteries used there, these estimates are also valid for original prospect theory. For separable prospect theory they may be somewhat different, but this does not affect the essence of the example here.

[7] Contrary to SB's suggestions, dependence on number of outcomes cannot solve this problem (§6).

[8] Such extensions depend too much on the particular discrete approximations chosen and they depend on π only through $\pi'(0)$.

between lotteries with no more than, say, four oucomes (Gonzalez & Wu 2022; Peterson et al. 2021; Starmer 1999; Wu, Zhang, & Abdellaoui 2005). However, rank-independent weighting does not work well beyond. Violations include Bernasconi (1994 p. 69), Blondel (2002 pp. 260-261), Edwards (1996 §III.B), Fennema & Wakker (1997), Fudenberg & Puri (2022 p. 422), L'Haridon (2009 p. 548), Levy (2008 all tasks), Loehman (1998 p. 293), Schneider & Lopes (1986 p. 546 first para), and Sonsino, Benzion, & Mador (2002 p. 946). Rank-independent weighting cannot be used in monotonic economic theories. Whereas SB (p. 1364) once acknowledged the problem of violation of stochastic dominance by rank-independent probability weighting, the rest of their paper only wrote positive about it and recommended a return to this old psychological theory. Section 6 returns to this point.

One of the main empirical findings for decision under risk concerns the overweighting of best and worst outcomes and the underweighting of intermediate outcomes (Fehr-Duda & Epper 2012; l'Haridon & Vieider 2019; Luce & Suppes 1965 §4.3; Starmer 2000). This fits well with rank dependence, but falsifies rank-independent probability weighting.

2.3. Non-identifiability of probability weighting and utility

SB aimed to measure probability weighting and utility. To do so, they only considered lotteries with one nonzero outcome in both their experiments. However, a joint power of probability weighting and utility then is unidentifiable. Thus, $\pi(p)u(x)$ is empirically indistinguishable from $\pi(p)^r u(x)^r$ for any $r > 0$ (Cohen & Jaffray 1988 Eqs. 5 and 7a). For this reason, Fehr-Duda & Epper (2012 p. 583) strongly advised against using only such stimuli.

SB claimed that

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

fits their data best in Experiment 1. However, $\left(\frac{p^{0.715}}{(p^{0.715}+(1-p)^{0.715})^{1/0.715}}\right)^{\frac{1}{0.941}} x$ fits their data exactly as well. In particular, the power family that they assume for utility can never rule out linear (or convex or concave) utility as best fitting. Similarly, in SB'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 exactly as well as their

claimed optimal fit. Hence, SB could not really identify utility and probability weighting.

The incorrect measurements complicate SB's claims on rank-independent probability weighting. We, therefore, do not discuss them further except in the last para of §6. We instead focus on SB's analyses of rank dependence henceforth.

3. Deterministic analysis of SB's experiments and linear utility

This section gives some preparatory mathematical definitions. In particular, it identifies an assumption of linear utility that SB took no account of, discussed further in later sections. We assume CPT, with the following evaluation of lotteries:

$$(p{:}\,X, q{:}\,Y, 1-p-q{:}\,Z) \;\rightarrow\; w_X u(X) + w_Y u(Y) + w_Z u(Z). \tag{3}$$

Here, u is as above, and w_X, w_Y, and w_Z are *decision weights*. Decision weights are nonnegative and add to 1. They are *rank-dependent*. For example, w_X depends on whether X is the best, middle, or worst outcome. We follow SB in using the term rank informally and in not expressing rank dependence in notation. A complete definition of general CPT, including reference dependence and a formalization and notation for ranks and rank dependence, is in Wakker (2010).

SB's first and main experiment concerned indifferences of the form

$$(p{:}\,X, q{:}\,Y, 1-p-q{:}\,Z) \sim (p{:}\,X, q{:}\,Y+m, 1-p-q{:}\,Z-k). \tag{4}$$

X, the common outcome, was varied across SB's main experiment, with m and k so small that the ranking of outcomes is the same for the two lotteries in Eq. 4 (i.e., "comonotonicity" is satisfied). Throughout, $Y = 24$, $Z = 18$, $m = 5$, and $q = 0.3$. Price lists were used to elicit k (called an *equalizing reduction* by SB) for three different values of p: $p = 0.1$, $p = 0.4$, and $p = 0.6$.

SB used seven values of X. In some instances, X was the best ranked outcome ($X = 34$, $X = 32$, or $X = 30$). In other cases, X, which we denote X', was ranked in the middle ($X' = 23$, $X' = 21$, or $X' = 19$). When X is ranked best, the weights of outcomes Y and $Y + m$ are denoted w_Y. When X is ranked middle, they are denoted $w_Y{}'$. We similarly write k'. Under linear utility, $\frac{w_Y}{w_Z} = \frac{k}{m}$ and $\frac{w_Y{}'}{w_Z} = \frac{k'}{m}$. The ratio

$$\frac{k'}{k} = \frac{w_Y'}{w_Y} \tag{5}$$

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

SB repeatedly claimed that they could have Eq. 5 for all differentiable utility functions. However, this claim was based on marginal rates of substitution involving infinitesimal changes m and k, which cannot be implemented empirically.[9] Empirically, we have to work with moderate outcome changes, and the following assumption:

ASSUMPTION 1 [linear utility for moderate outcome changes]. For outcome changes within a small interval $[A, B]$, utility is approximately linear.[10] □

SB's second experiment concerned indifferences of the form

$$(p{:}\,X, q{:}\,Y, 1-p-q{:}\,Z) \sim (p{:}\,X+m, q{:}\,Y-k, 1-p-q{:}\,Z-k). \tag{6}$$

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 SB again called an *equalizing reduction*. Under linear utility and Eq. 3,

$$w_X = \frac{k}{m+k}\,. \tag{7}$$

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

---

[9] SB's Footnote 13 even claimed 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 SB 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).

[10] For the studies considered in this comment, $[A, B] = [0{,}100]$ suffices. More precisely, for Eq. 5 and SB's first experiment, linearity of utility is used on all intervals $[\min\{Z-k, Z-k'\}, Z]$.

affect the decision weight in Eq. 7. SB again captured the change by the ratio $\frac{k'}{k}$.[11] Eq. 7 again uses Assumption 1. More precisely, it used linearity of utility in the intervals $[\min\{18-k, 18-k'\}, 18]$ and $[\min\{36-k, 36-k'\}, 36]$.

4. Small payoff changes: Ramsey's trifle problem in SB's experiment

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

> 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]

Samuelson (1960 Footnote 5) also pointed out this trifle problem. SB were not aware of it. In order to approximate infinitesimal changes with perfect linearity, they used very small payoff changes m, k. But these changes were too small to motivate subjects, making choice options almost identical. SB's subjects had to bother about trifles. Bear in mind that differences in outcomes, rather than outcomes themselves, give motivation for careful preferences.

Abdellaoui et al. (2020 §4) argued in detail that SB's experiments measured only heuristics and not preferences. In brief, there were too many choices with too small incentives. Smith (1982, "dominance"), Wilcox (1993), and many others warned against this. Ariely, Loewenstein, & Prelec (2001) warned against coherent arbitrariness (called the shaping hypothesis by Loomes, Starmer, & Sugden 2003) for such cases, where subjects develop coherent heuristics rather than coherent preferences. Online Appendix OA.1 (Wakker 2022c) shows mathematically that this will have happened in SB. Further, the layout in SB's Experiment 1 made *cancellation* (ignoring common outcomes as heuristic rather than as true preference)

[11] This result is not exact. 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. Probably, SB still used k'/k, or its log, as index in their analysis for this reason.

too salient. Weber & Kirsner (1997 top of p. 57) showed that Wakker, Erev, & Weber (1994) suffered from cancellation, explaining the absence of rank dependence there. Weber & Kirsner avoided it and then did find rank dependence. These papers used stimuli as in SB's first experiment (Eq. 4). Thus, SB's first experiment has replicated Wakker, Erev, & Weber's cancellation problem.

To avoid cancellation, SB (§5.3) carried out a second, smaller experiment, based on Eq. 6. Now there was no common outcome to be cancelled. The format in SB's second experiment has been used before but it never became popular because of its restriction to linear utility (§5). Hence, less is known about presence or absence of heuristics. Still, too many problems of SB's Experiment 1 remained. The complexity of the lotteries was worsened due to the absence of a common outcome. This augmented Ramsey's trifle problem. The layout of the stimuli (their Online Appendix, Figure 5) with, again, the same format for hundreds of choices over several pages, again induced coherent arbitrariness.

5. A measurement of equalizing reductions preceding SB

Diecidue, Wakker, & Zeelenberg (2007; DWZ henceforth) used the same equalizing reductions as in SB's second experiment, i.e., indifferences as in our Eq. 6; see their Eq. 3.2. Thus, following Weber & Kirsner (1997), they avoided the cancellation in Wakker, Erev, & Weber (1994). DWZ's primary purpose also was to test rank dependence (their main hypothesis on p. 185 and p. 195 2nd para) using quantitative parameter-free estimations rather than counting statistics, preceding SB. To avoid Ramsey's trifle problem, their outcome differences k, m were considerably larger than SB's, with m never below €20 (≈ $30 in 2020 value). To still use Assumption 1, these differences were not very large though. Using moderate outcome differences and linear utility (Assumption 1) is the only way in which SB's equalizing reductions can be used, and this is what DWZ did.

DWZ discussed and minimized all aforementioned experimental problems of SB (§4).[12] Their main findings supported rank dependence (DWZ p. 192 last para). DWZ also found violations of rank dependence. Decision weights sometimes changed even though ranks did not. SB's finding on rank dependence is reversed in the sense that decision weights did not change even if ranks did. This absence of rank dependence can be taken as a special case of CPT (expected utility), whereas DWZ's finding really falsifies CPT. Wu (1994), using different stimuli, found more extensive violations of rank dependence. The violations of rank dependence by DWZ, Wu, and others (§8) are more serious than SB's absence of rank dependence in the same way as the Allais paradox is a more serious violation of expected utility than risk neutrality.

Once it is understood that SB's equalizing reductions need linear utility, measuring indifferences and then calculating decision weights using linear utility, as SB (and DWZ) do, is not very original (l'Haridon & Vieider 2019 Eq. 3). Although linear utility can be defended for the moderate stakes used (DWZ p. 196; l'Haridon & Vieider 2019 p. 189), this method of DWZ, SB, and others never became very popular in economics. Most measurements of prospect theory allow for nonlinear utility.

---

[12] In the beginning of §3, DWZ explained that three-outcome lotteries 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. 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). DWZ developed their stimuli 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 emphasized that avoiding heuristics is desirable to increase statistical power (p. 188 last line; p. 195 3rd para). Two further differences between DWZ and SB are as follows. First, an improvement of one lottery was not compensated by worsening that same lottery elsewhere, but instead by improving the other lottery. This avoids confounds due to differences between improvements and worsenings. Second, DWZ considered, instead of risk with known probabilities, the more interesting context of uncertainty (ambiguity) with unknown probabilities. While this does not affect the theoretical working of rank dependence (DWZ, p. 185 ll. 7-8 and Wakker 2010 Figure 7.4.1 versus 10.4.1), it may, of course, impact differences in empirical findings.

6. Complexity aversion

SB (p. 1364) wrote:

> "probability weighting [rank-independent, as in in separable prospect theory] … implies violations of first-order stochastic dominance … This is a serious flaw …"

Nevertheless, the rest of the paper wrote positive about rank-independent probability weighting, returning to the displayed problem only at the end. SB then apparently suggested that "complexity aversion" (their term[13]) could resolve the problem. However, the central question then should have been whether complexity aversion can help *reduce* the problematic violations of stochastic dominance. SB did not address this question. Instead, they suggested that complexity aversion can *add* violations of stochastic dominance, considered desirable by SB. Online Appendix OA.3 (Wakker 2022c) does address the central question, with a negative answer: complexity aversion cannot help reduce violations of stochastic dominance. Thus, SB's suggested alternative for rank-dependent probability weighting does not solve the problem.

There are several other problems with SB's analysis of complexity aversion as a dependence on the number of outcomes. Many authors, not cited by SB, investigated such dependence before. Neilson (1992) proposed a formal model, but Humphrey (2001a) tested it unsuccessfully. Related models received some attention in psychology (Birnbaum 2008 p. 481; Krantz et al. 1971 Ch. 8; Luce 2000). Tversky & Kahneman (1992 p. 317) discussed the idea but were pessimistic:

> 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]

---

[13] SB (p. 1367 & p. 1402) wrote "a form of complexity aversion: people may prefer lotteries with fewer outcomes because they are easier to understand" and used this definition thoughout their paper.

I share his pessimism.

Another problem is that the prevailing empirical finding for gains, away from the certainty effect, is complexity seeking rather than the aversion claimed by SB, in studies not cited by SB. Those studies usually considered 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 require identical certainty equivalents, experiments find systematic differences. Here a pure effect of perceived number of outcomes occurs, clearer than the discontinuous changes of certainty equivalents considered by SB. Such effects are often called *event splitting effects* (or boundary effects, or violations of coalescing/collapsing).

Event splitting effects are special cases of attribute splitting effects (Weber, Eisenführ, & von Winterfeldt 1988), or the part-whole bias (Bateman et al. 1997), or, for uncertainty, the unpacking effect (Tversky & Koehler 1994). That is, splitting something up usually increases the total decision weight. This underlies several theories by Birnbaum cited below, who provided detailed analyses explaining why we mostly find complexity seeking but sometimes complexity aversion. A literature search gave:

- The following three papers report prevailing complexity aversion: Bernheim & Sprenger (2020), Huck & Weizsäcker (1999), Moffatt, Sitzia, & Zizzo (2015).
- The following eight papers report prevailing complexity seeking: Birnbaum (2005), Birnbaum (2007), Erev et al. (2017), 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).

Several of these references were given by Birnbaum (2008, p. 473), a paper cited by SB (Footnote 69) but not for these findings opposite to their complexity aversion. We conclude that the findings on complexity aversion are volatile, but the literature has documented more complexity seeking than aversion for gains.

The above discussion shows that dependency on number of outcomes is primarily driven by factors other than complexity perception, so that SB's term complexity aversion is a misnomer. Aversion to more comprehensive and fitting forms of complexity has been studied by Armantier & Treich (2016), Bruce & Johnson (1996), Kovarik, Levin, & Wang (2016), Mador, Sonsino, & Benzion (2000), and Sonsino,

Benzion, & Mador (2002). SB cited five papers on complexity aversion in their footnote 70, but they were all taken out of context.[14]

SB's reports of rank-independent nonlinear probability weighting are confounded by their incorrect measurements of probability weighting and utility. But probably their findings, if re-analyzed correctly, reflect event splitting effects rather than complexity aversion. Framing and preference reversal effects of this kind are known to be strong. However, they violate not only prospect theory but every transitive and monotonic economic theory. Humphrey (1995) pointed out that event splitting effects can be modeled using separable prospect theory with subadditive weighting, but Sonsino, Benzion, & Mador (2002 p. 946) found that it does not work well for this purpose.

7. SB's criticism of common statistical tests

SB (§2.3) criticized counting tests, i.e., statistical tests that compare numbers of violations of predictions, preference axioms in our case. Such statistcs are commonly used throughout decision theory and all empirical sciences. SB seeked to avoid crediting priority to much preceding literature this way. However, a fundamental problem in SB's discussion is that they did not know that *all* statistical tests, not only their own tests but also counting tests, are based on assumptions about probabilistic errors ("noise").[15] See:

> 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

[14] 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.

[15] Without it, p-values could not even be defined.

> violation frequencies are *necessarily*[16] higher for invalid axioms—is flawed. [italics and footnote added] [SB p. 1376]

SB's lack of understanding of this point also appears from SB (p. 1367 l. -3 and p. 1376 2nd para). Their claimed first problem (p. 1376) only shows the *existence* of an error model under which counting tests are incorrect. However, deviating error models exist for every statistical test (Greenland et al. (2016 p. 338 2nd column 1st para), including those used by SB. For example, if the errors in SB's indifference measurements are not constant or not stochastically independent, or extreme, then their claimed p-values and confidence intervals are not valid either. This is in fact plausible given that their test statistics were ratios of numbers close to 0 with big variances (Eqs. 5 and 7; see Online Appendix OA.1, Wakker 2022c). In general, the relevant question is not existence, but plausibility, of an error model deviating from the one required. SB, unaware of the latter concept, did not address this relevant question.

SB's second claimed difficulty concerns the example at the end of their Online Appendix B. It assumes stimuli for which CPT and expected utility have identical predictions. Then counting tests indeed have no power. But then, no statistical test does. This trivial example cannot serve as a criticism of counting tests or any other test. I was not able to understand SB's description of this example in their main text (p. 1376).[17] SB's claim to have refuted all counting tests, widely used in all empirical sciences, is completely unfounded. Therefore, SB should not have escaped from crediting priority to the many preceding papers that used counting statistics.

8. Further preceding falsifications of prospect theory

Given that prospect theory is the most tested theory of decision under risk, besides being most confirmed, it is also most violated. Wakker's (2022a) keyword

---

[16] Counting tests, as all statistics, do allow for deviating samples, be it with small probabilities not exceeeding a significance threshold.

[17] SB write: "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."

"PT falsified" gives 59 references (reproduced in Online Appendix OA.5, Wakker 2022c). We call special attention to a finding of Birnbaum & McIntosh (1996), not cited by SB. 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 is of special interest because it concerns lotteries of the same format as in Eq. 4, i.e., as in SB'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. SB quantitatively found no change in decision weighs. The aforementioned studies avoided heuristics and found a stronger deviation: a decrease, rather than increase, in weight. These violations (inconsistent rank dependence) are, again, more serious than SB's violations in the same way as the Allais paradox is a more serious violation of expected utility than risk neutrality.

The strongest counterexample to rank dependence that I am aware of is Machina's (2009) reflection example, confirmed empirically by l'Haridon & Placido (2010) and informally qualified as "brilliant" by Wakker (2022a).

9. A criticism of Bernheim, Royer, & Sprenger (2022)

Bernheim, Royer, & Sprenger (2022), BRS henceforth, redid part of SB's first experiment. They avoided cancellation and fatigue and improved stimulus explanations, as has been common in preceding studies (DWZ; Weber & Kirsner 1997). However, contrary to BRS's claims, they worsened rather than improved incentives. As explained above, *differences* in outcomes (m, k in BRS's notation), rather than outcomes themselves, provide incentives. Unfortunately, BRS mostly increased outcomes but not their differences, even though Abdellaoui et al. (2020 p. 2 l. 8, p.8 l. 9) had already warned against this. BRS's subjects still had to bother about trifles. In only one incentivized choice pair ("Condition 5") difference was increased ($m = 20$ iso of $m = 5$), though still below DWZ's minimal difference. Incentives were worsened because only few subjects were paid, $1/20^{th}$. Remarkably, BRS's Figure 1 did suggest rank dependence for this choice pair, but BRS gave no statistical analysis of it.

Apart from a trivial test of stochastic dominance (k^+ vs. k^-), all statistical conclusions in BRS were based on accepted null hypotheses. Every statistics textbook warns against this (Greenland et al., 2016, Misinterpretation 4). Power analyses would have been warranted.[18]

BRS ignored, and gave no counterarguments to, all of Abdellaoui et al.'s (2020) warnings other than those concerning experimental stimuli. Thus, BRS (Finding 2) continued to use the invalid measurements of probability weighting and utility, continued to claim novelty on equalizing reductions while not citing the preceding DWZ, continued to use an incorrect formula of original prospect theory, and continued to declare all counting statistics invalid (BRS end of §1). BRS did not address the inviability of their suggested alternatives of rank-independent weighting and complexity aversion. BRS's null hypotheses and Finding 1 did not falsify rank dependence but only showed neutrality. For real falsification, BRS should also have solved the many problems of their Finding 2 (which they did not reconsider).

10. Conclusion

Bernheim & Sprenger (2020, SB) and Bernheim, Royer, & Sprenger (2022) reported flawed experiments to falsify rank dependence and prospect theory. Had the experiments been carried out correctly, and found the falsifications claimed, then these findings still would not have been new. More serious violations of the same kind, taking away novelty (in particular, regarding equalizing reductions) have been reported before, as well as many other negative findings. Nevertheless, rank dependence and prospect theory remain the most popular risk theory today because of many more positive results and, importantly, absence of a viable alternative. In particular, SB's suggested alternatives of rank-independent weighting, better known as separable prospect theory, and "complexity aversion" (a misnomer), based on incorrect theoretical and empirical claims, have been known not to work for decades.

[18] Confidence intervals of within-subject differences, rather than the "between"-subject confidence intervals of BRS's Figure 1, would also have been useful.

It would have been surprising if SB had been the first to "properly" test rank dependence, 40 years after its introduction (Quiggin 1982; Schmeidler 1982), 30 years after its incorporation into prospect theory (Tversky & Kahneman 1992), 20 years after its shared prize in memory of Nobel in 2002, and after thousands of applications.

References

Abdellaoui, Mohammed, Chen Li, Peter P. Wakker, & George Wu (2020) "A Defense of Prospect Theory in Bernheim & Sprenger's Experiment," mimeo; https://personal.eur.nl/wakker/pdf/abd.li.wak.wu_bernh.sp.pdf

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

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

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

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.

Bernasconi, Michele (1994) "Nonlinear Preference and Two-stage Lotteries: Theories and Evidence," *Economic Journal* 104, 54–70.

Bernheim, B. Douglas, Rebecca Royer, & Charles Sprenger (2022) "Robustness of Rank Independence in Risky Choice," *AEA Papers and Proceedings* 112: 415-420.

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. (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.

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.

Blondel, Serge (2002) "Testing Theories of Choice under Risk: Estimation of Individual Functionals," *Journal of Risk and Uncertainty* 24, 251–265.

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

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

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.

Dean, Mark & Pietro Ortoleva (2017) "Allais, Ellsberg, and Preferences for Hedging," *Theoretical Economics* 12, 377–424.

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.

Edwards, Kimberley D. (1996) "Prospect Theory: A Literature Review," *International Review of Financial Analysis* 5, 18–38.

Erev, Ido, Eyal Ert, Ori Plonsky, Doron Cohen, & Oded Cohen (2017) "From Anomalies to Forecasts: Toward a Descriptive Model of Decisions Under Risk, Under Ambiguity, and From Experience," *Psychological Review* 124, 369–409.

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.

Fishburn, Peter C. (1978) "On Handa's "New Theory of Cardinal Utility" and the Maximization of Expected Return," *Journal of Political Economy* 86, 321–324.

Fudenberg, Drew & Indira Puri (2022) "Simplicity and Probability Weighting in Choice under Risk," *AEA Papers and Proceedings* 112, 421–425.

Gonzalez, Richard & George Wu (2022) "Composition Rules in Original and Cumulative Prospect Theory," *Theory and Decision* 92, 647–675.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, & Douglas G. Altman (2016) "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology* 31, 337–350.

Hirshman, Samuel D. & George Wu (2022) "Tests of Rank-Dependent Probability Weighting in Risky Choice," in preparation.

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.

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

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.

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

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

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 (2009) "Behavior in the Loss Domain: An Experiment Using the Probability Trade-Off Consistency Condition," *Journal of Economic Psychology* 30, 540–551.

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.

Levy, Haim (2008) "First Degree Stochastic Dominance Violations: Decision Weights and Bounded Rationality," *Economic Journal* 118, 759–774.

Loehman, Edna (1998) "Testing Risk Aversion and Nonexpected Utility Theories," *Journal of Economic Behavior and Organization* 33, 285–302.

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.

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.

Murphy, Ryan O., & Robert H.W. ten Brincke (2018) "Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates," *Management Science* 64, 308–326.

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

Nilsson, Håkan, Jörg Rieskamp, Eric-Jan Wagenmakers (2020) "Commentary: Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory," *Journal of Mathematical Psychology* 98, 102429.

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.

Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, & Thomas L. Griffiths (2021) "Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making," *Science* 372, 1209–1214.

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

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. (1960) “The St. Petersburg Paradox as a Divergent Double Limit,” *International Economic Review* 1, 31–37.

Schmeidler, David (1982) “Subjective Probability without Additivity,” Foerder Institut of Economic Research, Tel Aviv University, Tel Aviv, Israel.

Schmeidler, David (1989) “Subjective Probability and Expected Utility without Additivity,” *Econometrica* 57, 571–587.

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

Schneider, Sandra L. & Lola L. Lopes (1986) “Reflection in Preferences under Risk: Who and when May Suggest why,” *Journal of Experimental Psychology: Human Perception and Performance* 12, 535–548.

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

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 (1999) “Cycling with Rules of Thumb: An Experimental Test for a New Form of Non-Transitive Behavior,” *Theory and Decision* 46, 141–158.

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.

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

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

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

Wakker, Peter P. (2022a), "Annotated Bibliography," http://personal.eur.nl/wakker/refs/webrfrncs.docx

Wakker, Peter P. (2022b) "The Correct Formula of 1979 Prospect Theory for Multiple Outcomes," *Theory and Decision*, forthcoming.

Wakker, Peter P. (2022c) "Online Appendix of: A criticism of Bernheim & Sprenger's (2020) Tests of Rank Dependence," http://personal.eur.nl/wakker/pdfspubld/sb.criticism.onl.appndx.pdf

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

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

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.

Wu, George, Jiao Zhang, & Mohammed Abdellaoui (2005) "Testing Prospect Theories Using Tradeoff Consistency," *Journal of Risk and Uncertainty* 30, 107–131.