A criticism of Bernheim & Sprenger (2023)[1]

Peter P. Wakker

Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands, Wakker@ese.eur.nl

February 2024

ABSTRACT

To avoid admitting mistakes in their preceding works pointed out by Wakker (2023 JBEE), Bernheim & Sprenger (2023 JBEE) use fallacies and miscitations, most of them easy to see through.

[1] The Journal of Behavioral and Experimental Economics announced beforehand that it will not publish any follow-up discussion, including this note. This note will serve as part of Wakker, Peter P. "Annotated Bibliography", http://personal.eur.nl/wakker/refs/webrfrncs.docx, March 2024 and later.

§1 Introduction

The core of this paper, §3 (6 working-paper pages), shows that Bernheim & Sprenger (2023; 21 journal pages), used fallacies and miscitations to cover up preceding mistakes in Bernheim & Sprenger (2020) and Bernheim, Royer, & Sprenger (2022), pointed out by Wakker (2023a).

ABBREVIATIONS (easy to remember via year/subscript)
SB_0: Bernheim & Sprenger (2020)
SB_2: Bernheim, Royer, & Sprenger (2022)
SB_3: Bernheim & Sprenger (2023)
SB: SB_0, SB_2, & SB_3
W_0: Abdellaoui et al. (2020) (subsumed by W_3)
W_3: Wakker (2023a)
§: sections in this note
§: sections in other papers
SPT: separable prospect theory $\left(\sum w(p_j)u(x_j)\right.$: "rank-independent weighting"$\left.\right)$
CPT: Tversky & Kahneman's (1992) cumulative prospect theory (rank-dependent)

For self-containedness, §2 repeats some mistakes by SB pointed out by W_3. The subsections in §3 are in decreasing order of transparency-of-mistake. SB's refusal to admit an elementary well-known mistake, in §3.1, is telling. §3.2-§3.11 list many further fallacies and miscitations in SB_3, all elementary. The main claim of SB_3 is that W_3 would have ignored, and/or misunderstood, SB_0's Method 2.[2] But W_3 did understand it (§3.12 and §3.13). There are two pleas of guilt, in §3.12 and Footnote 17. After three technical appendixes, Appendix D links SB_3's claims to §3 in order of their appearances in SB_3. It does not bring new points besides small details in small font, making this paper concise through §3.

[2] SB_3 p. 6 §2.1.2 last para: "This fundamental error … infects most of the Wakker commentaries' critiques"; etc.

§2 Mistakes of SB pointed out by W_3

SBM1. SB's data reflected heuristics, due to insufficient stakes and complex stimuli.

SBM2. SB claimed to identify unidentifiable utility and probability weighting.

SBM3. SB needed approximately linear utility but claimed general validity.

SBM4. SB proposed no viable alternative to CPT because:

SBM4a. SB's rank-independent weighting has no sound revealed-preference meaning, serving in no sound decision theory (SPT is not sound).

SBM4b. SB's complexity aversion is not part of a useful, sufficiently specified theory. It is empirically even in the wrong direction.

SBM5. SB_0 mislabeled SPT as 1979 prospect theory, adding to the confusion.

SBM6. SB improperly claimed priorities.

SBM7. SB's claims on statistics—to escape crediting priority—were incorrect and revealed elementary lacks of understanding.

§3 Criticizing SB's mistakes

SB_3 often refer to Appendices, but those are not available at the journal website. Similarly, supplemental materials such as the data set are not available at SB_2's journal website.[3] These absences prohibit verifications of SB's claims.

§3.1. Identifying the unidentifiable

It is unbelievable that SB_3 maintain that SB_0 (§3.2) identified utility u and probability weighting π in $\pi(p)^r u(x)^r$ from lotteries with only one nonzero outcome[4], even though this is a well-known mistake: the joint power r of utility and probability weighting then is even mathematically[5] unidentifiable (W_3 §2.3). Fehr-Duda & Epper (2012 p. 583 2nd para) warned against this mistake. SB_3 (p. 7 3rd para) counter that the power of utility, 0.941, randomly produced by their software, is close to linearity, which is empirically plausible (see also SB_3 §2.2 ℓ. 5; etc.). However, that 5 is empirically close to 4 cannot justify the mathematical claim $2 + 2 = 5$.

This case transparently shows that SB will not admit any mistake, even if elementary and for everyone to see. More will come.

§3.2. Covering up a stakes-mistake by miscitations

W_0 pointed out that SB_0's stakes m, k, denoting eventwise *differences* between outcomes of different lotteries, were too small. Unfortunately, SB_2 did not understand and mostly increased the wrong amounts: absolute outcomes but not their differences (W_3 §9).[6] Accordingly, SB_2 only found H_0s. To escape from admitting their mistake, SB_3 (p. 18) erroneously claim that W_0 had not properly pointed out the stakes problem:

> Wakker [W_3] *now* asserts that only [I never claimed "only"] the difference between payoffs *Y* and *Y* + *m* … are relevant for assessing stakes. ALW&W's [W_0] claims about inadequate stakes invoked no such distinction, which is why BR&S [SB_2] did not focus more narrowly on inflating the value of *m*. Wakker's emphasis *shifted* after the publication of BR&S [SB_2]. … this claim is an *ex post* rationalization. [italics added]

To the contrary, W_0 was crystal clear about m and k and W_3 did not shift emphasis. W_0 (p. 2) and W_3 (p. 1) already announced that SB_0's stake problem is that outcome

---

[3] Both absences still held on 25 February 2024.

[4] See SB_3 p. 2 Column 1 $\ell\ell$-12/-9, Footnote 9, and p. 14 first column $\ell\ell$. -2/-1; etc.

[5] That is, if perfect data with no noise.

[6] Except in one incentivized choice situation.

differences are too small, both italicizing “differences”. W_0 (p. 8) then unambiguously explained the crucial role of those differences/changes m, k:

> BS [SB_0] took payoff *changes* m, k that are very small. But these changes became too small to motivate subjects. … questions … which involved *nearly-identical* lotteries [italics added]

W_3 (§9) wrote, in full consistency with W_0:

> Unfortunately, BRS mostly increased outcomes but not their differences, even though Abdellaoui et al. (2020 p. 2 ℓ. 8, p. 8 ℓ. 9) [W_0] had warned against this.

SB_3’s claim that W_3 shifted emphasis is completely off.

Further, SB_3 (p. 18 Columnn 2) claim, out of the blue and incorrectly:

> “The relevant payoff differences in the B&S [SB_0] and DW&Z [Diecidue et al.] experiments are in fact similar, and the differences in BR&S [SB_2] are larger when m = \$20 [maximal m].”

In reality, Diecidue et al.’s (2007) average stakes were almost three times higher than those of SB_2 and the maximal stakes were the same (Appendix C).

SB_3 miscite W_0/W_3, and make incorrect unfounded claims, to cover up mistakes.

§3.3. *Complexity aversion: wrong empirical direction and terminological mistake*

It is amazing that SB_3 maintain their empirical claims about complexity aversion[7], even though W_3’s (§6) literature survey provided opposite evidence.[8] Further, SB_3 (§7.2 last para etc.) throughout misunderstand W_3’s (in fact, their own in SB_0!) term complexity aversion. SB_0 (p. 1367 4th para) take complexity aversion in a specific sense: aversion to lotteries with many outcomes. Only to avoid confusion, W_3 reluctantly followed SB_0’s unfortunate terminology.[9] He criticized this specific complexity aversion. SB_3 erroneouly criticize W_3 for criticizing general, rather than specific, complexity aversion (SB_3 p. 4 Claim #14 $\ell\ell$. 7-8, “generally”, etc.).

SB_3 ignore counterevidence to their complexity aversion and miscite.

§3.4. *Denying something as undeniable as a definition: 1979 Prospect Theory*

SB_0 mislabeled SPT for 1979 prospect theory (OPT; see SBM5), as pointed out by W_0 (after their Eq. 2). Wakker (2023b) published the correct definition of OPT for multiple outcomes, that is, how Kahneman and Tversky defined it, documenting it beyond dispute. To counter Wakker’s (p. 186 2nd para) argument that Kahneman and Tversky explicitly provided OPT’s probability weighting formula in the 1975 working paper version of their 1979 paper, SB_3 completely unfoundedly and erroneously claim

---

[7] SB_3 p .4: “the profession has *not* discarded simple aversion to the number of outcomes” (italics from original).

[8] The purest tests (W_3 §6) are taken as “different”, and are then ignored, by SB_3 (p. 21 3rd para).

[9] Explained in W_3, Footnote 12, but reluctantly so, calling it a misnomer (W_3 p .4 ℓ.-4).

that Kahneman and Tversky had “jettisoned” their 1975 weighting formula (SB_3 p. 19 para –4). To counter Wakker’s citation of Kahneman and Tversky’s (1979) verbal statement of the formula, SB_3 miscite Wakker (2023b); see my displayed “Analysis of Subtle Miscitation” in Appendix D referring to SB_3’s p. 19 para –3. Further, SB_3 ignore Wakker’s (2003b end of §3) citations of two later texts where Kahneman and Tversky confirmed OPT and distanced themselves from SPT. I add here that SPT violates Kahneman & Tversky’s (1979) Eq. 2, so that it cannot even qualify as a possible generalization, contrary to many claims by SB (e.g., SB3 p. 19 para –2).

SB_3 also erroneously claim that I recommended OPT as useful (SB p.2 Claim#1 middle; etc). To the contrary, W_3 (§2.2) pointed out that his example of bad (also descriptively) violation of stochastic dominance applies to both SPT and OPT. Both are unsound. Thus, OPT’s “eggregious” implications, emphasized by SB_3 (pp. 19-20), had been pointed out before by W_3, a point missed by SB_3.

SB_3 should have simply admitted their mislabeling of SPT rather than using deliberate miscitations to dispute something as indisputable as a definition.

§3.5. No data file provided and misplaced accusation of cherry picking

SB_3 (p.3, (Claim #8) ℓ.-5 & p.18 Column 2 2nd & 3rd para) accuse W_3 of cherry picking. However, W_3 discussed SB_2’s “Condition 5” only because it is the only incentivized choice with nontrivial stakes in all SB’s experiments, and W_3 (§9 end of first para) conjectured significant rank dependence there. Unfortunately, SB_3 still do not provide the statistics, and SB_2 did not provide the data set, so that it still can’t be verified. BB_3’s defensive term “cherry picking” does suggest statistical significance there.

§3.6. Misunderstanding elementary statistical principles

SB_3 (§3) continue in their, failing, attempts to criticize well-established counting statistics. They there fill another three journal pages with, apparently, scenarios to give Type I/II errors. W_3 (§7) criticized SB_0 for misunderstandings of statistics.

One illustration of SB’s escape attempts: SB_0 did not know that counting statistics, like every statistical analysis, assumes an underlying error (stochastic) model (W_3 §7). How else get p-values? SB_3 now avoid such explicit claims.[10] They, implicitly, suggest that such error theories, described in 100s of textbooks, are not

---

[10] Note SB_3’s (p .11 ℓ. 5) ambiguous writing to duck the question.

"standard" (SB_3 p.4 2nd para ℓ.–10; p. 9 Column 2 ℓ. 2; p. 10 penultimate para of §3,3,1; §3.3.3 ℓ. 5), not "reasonable" (SB_3 §3.3 ℓ. 3), and not "natural" (SB_3 §3.3.2 ℓ. -4). They qualify the many preceding studies as "suffer from important conceptual problems", with "endemic deficiencies" and "severe design flaws" (SB_3 §3 first para; §3.2 ℓ. -5; §5 last para).

SB_3 should have admitted priority of studies using standard counting statistics.

§3.7. No novelty in falsification of prospect theory

There have been numerous tests and falsifications of rank dependence and other aspects of prospect theory, unknown to SB or miscited[11]. W_3 (§10) explained the naivety of SB to think that, almost 30 years after Tversky & Kahneman (1992) with thousands of citations and a shared memorial Nobel prize, SB_0 could have been the first to "properly" test rank dependence of CPT and Quiggin (1982).

SB should have admitted prior falsifications of CPT. Instead, they erroneously try to dismiss them all (SB_3 §5 last para; end of our §3.6).

§3.8. Continued refusal to admit assumption of approximately linear utility

As did Decidue et al. (2007), SB should have admitted their (reasonable: W_3 §5 last five lines) assumption of approximately linear utility. The more so as they needed larger stakes than used (§3.2.), too large to call infinitesimal. Instead, SB_3 continue to sometimes erroneously claim full general validity (p. 5 ℓ. 6 etc.) but at other times claim reasonable approximations (p. 3 Claim #3 etc.). I add here: of linear utility!

§3.9. Ignoring priorities of Diecidue et al. (2007)

Diecidue et al. (2007) (DWZ) is *very* close to SB_0 (see W_0 pp. 11-13). DWZ also measured equalizing reductions, using exactly the same format as SB_0 in their supplemental experiment, which is close to the tasks in SB_0's main (first) experiment. DWZ did not involve counting statistics either, also avoided the cancellation heuristic (crediting Weber & Kirsner, 1997), and also used their quantitative measurements to test rank dependence. DWZ's degeneracy tests directly tested additivity of decision weights, equivalent to linearity.[12] SB_0's Method 2 did so only indirectly by also

[11] E.g., SB_0 cited Birnbaum (2008) and Weber & Kirsner (1997) but ignored their opposite findings.

[12] By Cauchy's equation (Aczèl 1966). Whereas the weighting function is nonadditive, decision weights are still additive under rank dependence or, equivalently, they always add to 1 for a given prospect. Nonadditivity of decision weights, as under SPT, necessarily implies violations of

involving equalities obtained from their Method 1. Appendix B explains the latter point, with a simple numerical example giving the gist.

SB_3 (p. 4 1st para etc.) do not discuss the big overlap just described, but only enlarge details of difference, that: (1) they do not need linear utility, which is incorrect; (2) they assume probabilities known whereas in DWZ those are unknown; however, this only makes DWZ more interesting and innovative[13]; (3) their combination of Method 2 and 1 is new; however, it is only a roundabout way of testing additivity of decision weights done more directly in DWZ's degeneracy tests (§3.13).

SB_2 should have cited the very close DWZ pointed out by W_0 but did not do so, criticized by W_3. SB_3 should have admitted DWZ's relatedness but still do not do so.

§3.10. Providing no viable alternative

SB should at least have suggested a viable direction of improvement of CPT. Despite SB_3's partial retractions (p. 4 Claims #13 and #14), SB_0 clearly propagated rank-independent weighting + complexity aversion.[14] (They surely suggested no other alternative!) But those have been known not to be viable (§3.3 & §3.4).

SB did not suggest any viable alternative to rank dependence and CPT.

§3.11. Putting up strawmen on heuristics

Whenever W_3 suggested a heuristic, SB_3 put up the strawman that that heuristic would apply to all subjects in all situations, with no other determinants of choice (p. 13 para on Columns 1-2; etc.). SB_3 then take every other significant effect as disproving the heuristic.[15] Thus, SB_3 duck the relevant question of plausibility of the heuristics.

In general, it is impossible, surely in retrospect, to prove that heuristics occurred. One can only argue for their plausibility. For example, W_0 once suggested a blending heuristic as plausible among others, and W_3 did not mention it, but SB_3 discuss it extensively (§6.2.7 and many other places) as if assumed universally. Contrary to many claims by SB_3 (p. 3 Claim #6 $\ell\ell$. 5-7 etc.), findings in the supplemental

---

monotonicity. The latter is crucial in the derivation of rank-dependent utility. See p .487 2nd para in Quiggin & Wakker (1994), which corrected Quiggin (1982).

[13] DWZ were the first to measure Schmeidler's (1989) nonadditive event weighting function quantitatively.

[14] SBo (p. 1367):"We hypothesize that the observed behavior results from a combination of standard PT [SPT] and a form of complexity aversion: people may prefer lotteries with fewer outcomes because they are easier to understand." SBo (p. 1402): "promising possibility is that the observed behavior reflects a combination of standard PT and a form of complexity aversion."

[15] P .3 middle of Column 1: "These patterns are inconsistent with any simple heuristic"; etc.

experiments where no cancellation can occur do not disprove cancellation (or other heuristics) in SB_0's main experiment. The accepted H_0s in SB_2 in fact prove nothing.

SB_3 should have admitted that heuristics were plausible in SB_0 and SB_2.

§3.12. Because SB_0's Method 2 was plagued by mistakes, W_3 did not sort them out but focused on the relatively cleaner Method 1

OBSERVATION 1. SB_0's Method 2 did not use the (never admitted) mistaken measurements of their §3.2. □

SB_3 put very central that W_3 did not make this Observation, and I *plead guilty* here. I feel justified because of the other mistakes (§2) remaining. It is impossible for readers to sort out which mistakes do or do not play a role where. In particular, it is impossible for readers to sort out Observation 1 from SB_0 (Appendix A below). W_3 mostly focused on SB_0's Method 1 because SBM2 and SBM4a play no role there, so that there may at least be sensible concepts involved with potential interest.

SB_3 should have admitted the mistakes in SB_0, rather than at length ruling out only one (never-admitted) mistake from one analysis while the other mistakes remain.

§3.13. W_3 did not misunderstand SB_0's Method 2

SB_3 put very central that W_3 would not have understood their Method 2. But W_3 did understand the method and criticized it. To prepare, SB_0's Method 1 found that decision weights of events do not change if ranks change. SB_0's Method 2 found "nonlinear decision weighting" if probabilities change, but used Method 1 in doing so.

W_3 (§6 last para) did point out that a correct analysis of SB_0's Method 2 (≠ SB_0's confused analysis!) would ("probably") reveal violations of CPT. *W_3 understood and acknowledged that!* The violations are not new but concern event/attribute-splitting in a roundabout manner. W_3 only mentioned it briefly but Appendix B below elaborates on it. W_3 cited two studies pointing out that such event splitting could also be accommodated by SPT, as SB_0 did. W_3 also cited Sonsino, Benzion, & Mador (2002) which explicitly criticized such modeling through SPT, supporting my SBM4a. Birnbaum's (2008) RAM & TAX models are the most advanced ones for such phenomena, but their refined psychological dependence on framing and violations of monotonicity are not suited for economic applications.

Appendix A. W3 did not make Observation 1

SB_0's quantitative estimates used in their Method 2 were obscured by their many mistakes and confusions (§2). It is impossible for readers to sort out which mistakes were involved where. SB_3 go to great lengths to repeat SB_0's analyses, apparently to show Observation 1 (SB_3 p. 2 ℓ.-4/-3; etc.). But the other mistakes (§2) remain!

Although SB_0 never said that they used §3.2 in their Method 2, as emphasized by SB_3 (p. 2 Column 1 ℓ. -7), SB_0 neither ever said that they didn't, contrary to many suggestions by SB_3 (p. 2 Column 1 ℓ. -9/-6; p. 6 2nd full para ℓ. 4-5; §2.2 ℓ. 10). That is, they never stated Observation 1. The Observation is almost impossible for readers to sort out from SB_0, and even surprising (why not use §3.2 to refine the measurements of probability weights?), contrary to many claims in SB_3. SB_3 were first to state Observation 1 explicitly. The observation is not important because of the other mistakes. I, therefore, feel justified in pleading guilty on W_3 not having sorted out Observation 1 and on W_3 mostly focusing on SB_0's Method 1.

Appendix B. SB_0's Method 2 as event splitting

SB_0's Method 2 involves known ways of violating CPT, due to what are called collapse effects or event splitting effects, as first illustrated here through a simple example. Assume, using SB_0's notation, probabilities $p = q = r = 1/3$, where p has the best outcome, q the second best, and r the worst. SB_0's analysis denoted decision weights by w and used proportional decision weights such as $\frac{w(p)}{p}$, $\frac{w(q)}{q}$, and $\frac{w(r)}{r}$, with the same probabilities but in different ranking positions. Under rank dependence the fractions can be different. However, SB_0's Finding 1 (of Method 1) is that they are not. CPT then implies that the proportional decision weight $\frac{w(p+q)}{p+q}$, where we now change the probability p into $p + q$ but do not change its best ranking position, must also be the same. SB_0's Finding 2 (of Method 2) is that it is not.

The nonlinearities w.r.t. size of probability found by SB_0's Method 2, in combination with Method 1, are equivalent to violations of additivity, $w(p) + w(q) \neq w(p + q)$ (Footnote 12). Thus, if two events have different outcomes, their total weight is different than if they have the same outcome. This is also a form of event and attribute splitting (the term used by W_3 p. 5 2nd para): splitting up changes the total. It necessarily leads to violations of monotonicity. The violations of CPT in SB_0's Method 2 are algebraic rewritings of this phenomenon. The involvement of their Method 1 only makes it more roundabout. Diecidue et al. (2007) tested such violations more directly through what they called degeneracy effects: decision weights when outcomes collapse versus when they do not. They thus found violations of CPT, as did SB_0's Method 2, in two of their six tests.

Appendix C. Diecidue et al.'s (2007) stakes were almost three times higher than those of SB_2

Preparatory calculations: the *maximal* m of SB_2 was \$20 (their "Condition 5"). The *minimal* B of Diecidue et al. (2007), DWZ henceforth (SB_3's Panel A chose a minimal B; I will refer to that panel) was $33 - 13$ = DFL 20. As SB_3 point out correctly, the mathematical analog of their m is not DWZ's B, but DWZ's $B - xk$ which is $33 - 19$ = DFL 14 in SB_3's Panel A, right matrix (the most unfavorable matrix there for me). It can be argued though that B, rather than $B - xk$, was made salient to DWZ's subjects, but let me nevertheless use $B - xk$ = DFL 14 henceforth, again the most unfavorable case for me. DFL 14 in 2001 (DWZ's guilders must have been implemented *before* 2002) is \$10 in 2021, the year before SB_2 appeared. The average B of DWZ (Table 2) was not the aforementioned minimal DFL 20, but DFL 26.66, being \$13.33 in 2021. SB_2 used their maximal $m = 20$ in *only one* incentivized indifference measurement, their Condition 5. In all other incentivized measurements of indifferences, SB_2 kept $m = 5$, as in SB_0. In stark contrast to SB_3's claims (§3.2), the conclusion should be:

> The incentivized relevant payoff differences in SB_2 are inferior to DWZ by a factor of almost 3 (13.33/5), except one single indifference measurement (SB_2's Condition 5), their maximum, which is equal to DWZ's maximum (B = DFL 40 in 2001 $\Leftrightarrow$ m = \$20 in 2021).

SB_3's Footnote 45 compares the random incentive implementations of SB_2 and DWZ. I next argue that DWZ's implementation of random incentivization is preferable. SB_2's only increased indifference measurement, Condition 5 (with their maximal stake \$20 in 2021), put central by them, has an implementation probability 1-in-180. For DWZ's maximal stake of B = Dfl 40 (also m =\$20 in 2021) it was 1-in-

220, which is comparable. SB_2 have two measured indifferences, with low stake $5, with increased implementation probability 1-in-45, considerably higher than DWZ's which was always 1-in-220. However, to achieve high implementation probabilities, SB_2 paid a heavy price. First, they implemented every of their four incentivized indifference measurements[16], increasing implementation probabilities by a factor four but losing incentive compatibility due to portfolio and hedging effects. DWZ implemented only one of 22, avoiding those violations. Second, SB_2 reduced the number of tasks by using a version of the Becker-DeGroot-Marschak mechanism, controversial for its complexity. DWZ used the well-established choice lists. To conclude, DWZ's implementation of random incentivization is preferable to SB_2's.

Appendix D. Responding to SB_3's claims in order of appearance

This Appendix replies to SB_3's criticisms in order of appearance, usually referring to §3 for relevant replies, and with sometime details added in small font. Below, Cm_1 abbreviates 1st column, and Cm_2 abbreviates 2nd column.

P. 2 $\ell\ell$. 6-7 ("Inexplicably … text."): §3.12 & §3.13 (Method 2 understood)
P. 2 4th para $\ell\ell$. 4-5 ("Contrary … CPT"): §3.12 & §3.13 (Method 2 understood)
P. 2 5th para $\ell\ell$. 1-2 ("Critically … B&S"): §3.12 & §3.13 (Method 2 understood)
P. 2 Cm_1 last para: §3.12 & §3.13 (Method 2 understood)
P. 2 Cm_1 last para $\ell\ell$. 3-6 ("Specifically … (1992)": §3.1 (unidentifiability)
P. 2 Cm_1 last para ℓ.-8/-6 ("we made it clear … that they [their §3.2] are not part of our formal analysis": not true. SB_0 never stated the above Proposition 1.
P. 2 Cm_1 last para ℓ.-7/-6 ("they are not even mentioned in Section 2": there is a big difference between not saying that and saying that not (Appendix A).
P. 2 Cm_1 last para ℓ.-5 ("email exchange"): guilty plea in §3.12 (mistakes not sorted out)
P. 2 Claim #1: §3.4 (incorrect PT formula). I add here Kahneman & Tversky (1979, p. 274: "seggregation"), again distinguishing OPT from SPT.
P. 2 Claim #1 ℓ.14 ("Wakker recommends"): Wakker never "recommended" it (§3.4).
P. 2 Claim #1 ℓ.-7: SB_3 mention the mislabeling of SPT, but without admitting (§3.4).
P. 2 (Claim #1 $\ell\ell$.-6-end): for SB_3's implausible conjecture of all lotteries mixed, burden of evidence is with them.

---

[16] Note that SB_2's experiment has only four incentivized indifference measurements, whereas DWZ had 22, greatly facilitating SB_2's implementation probabilities.

Pp. 2-3 (Claim #2): §3.1 (unidentifiability), §3.12 (mistakes not sorted out), §3.13 (Method 2 understood), §3.8 (linear U).

P. 2 Claim #2 ℓ. 3: Wakker was not mistaken because he referred to identifying not only probability weighting but also utility, which is only BS_0's §3.2 and not their Method 2. Same mistake is on SB_3 p. 7 ℓ. 6.

P. 2 ℓ. -2 till p. 3 ℓ. 2 ("Wakker contradicts his own … "): §3.1 (unidentifiability),

P. 3 (Claim #3): §3.8 (linear U)

P. 3 (Claim #4): §3.11 (heuristics). Contrary to SB_3's long text, W_3 does not claim that SB_0's tasks are too difficult, or too numerous, in an absolute sense, but they are so jointly, given also the overly small stakes.

P. 3 (Claim #4 ℓ.-11): SB_3 write, misleadingly: "His claim that our subjects made "hundreds of choices" improperly counts each line of a price list as a separate decision. The main B&S experiment involved 28 elicitations, not hundreds." However, W_0 (p. 8) wrote: "subjects completed 28 price lists … subjects answered 980 (21 × 38 + 7 × 26) questions."

P. 3 (Claim #5): §3.11 (heuristics). OA.1 *proves* that true preferences cannot have the precision found by SB_0. Contrary to SB_3's claim (also §3 last para), this *proves* that SB_0 can't have measured true preferences.

P. 3 (Claim #6): §3.11 (heuristics)

P. 3 (Claim #7): §3.12 (mistakes not sorted out) & §3.13 (Method 2 understood).

P. 3 (Claim #8): §3.2 (stake-size). I add here that Kahneman & Tversky (1979 p. 275 4th para) pointed out that small differences between prospects may be ignored.

P. 3 (Claim #8) ℓ.-5 ("cherry-picking"): see *§3.5*.

P. 3 (Claim #8): "he misstates the stakes" See comment below to p. 18 Cm_2 top.

P. 3 (Claim #9): In an email to Hirshman and Wu of 20 November 2023 I explained that this misunderstanding between them and me could not have been avoided.

P. 4 (Claim #10): §3.9 (Novelty of equalizing reductions)

P. 4 ℓℓ. 14-15 "no counterparts": DWZ's degeneracy tests are more direct (§3.9).

P. 4 (Claim #11): §3.7 for novelty, §3.6 for statistics, and §3.12 for Method 2

P. 4 (Claim #11) ℓℓ.-9/-8: "Wakker was not able to understand" (W_3's §7 4th para): that was a polite way of saying that SB_0's example is completely nonsensical. As should be clear from W_3's writing there. SB_3, not familiar with such a mode of expression, take it literally. They also do so on p. 9 Cm_2 ℓ.9: and in their Footnote 17.

P. 4 (Claim #12): §3.7 (preceding falsifications of PT)

P. 4 (Claims #13 & #14): §3.10 on alternatives & §3.3 on complexity aversion

P. 4 (Claim #15): Wide consensus is that CPT is currently best but has problems.

Pp. 4-5 §2.1.1 repeats large parts of SB_0's analysis. Probably just to show that SBM2 plays no role. But the other mistakes remain (§3.12).

P. 5 ℓ. 6 "marginal utilities cancel": see SBM3. Not true. Marginal utilities only cancel approximately, if U is approximately linear.

P. 6 §2.1.2: §3.12 & §3.13 (Method 2 understood)

P. 6 §2.1.3: §3.8 (linear U). The good approximations claimed are of linear utility!

Pp. 6-7 §2.2: §3.1 (unidentifiability)

P. 7 Footnote 9 continues to erroneously claim the identification (§3.1).

P. 7 ℓ. 3-: SB_3 write there:

> "he [Wakker] … writing … "SB [SB_0] aimed to measure probability weighting and utility. *To do so, they only considered lotteries with one nonzero outcome in both their experiments*" ... These statements are simply false" [italics not in W_3 but added by SB_3]

Wakker's cited statements are correct and not "false"! They are correct because they also refer to utility estimation, which SB_0 only tried to do (incorrectly) in both their experiments (in §3.2 and in §5.3's line on pp. 1396-1397) by using lotteries with only one nonzero outcome, and not in their Method 2. SB_3's mistake on p. 7 is to consider only probability weighting and to miss the utility part.

P. 7 3rd para: §3.8 on linear utility. See also comment on their p. 2 ℓ. -2 till p. 3 ℓ. 2.

P. 7-8 §2.3: §2.3.1 & §2.3.2 repeat SB_0 & SB_2.

P. 8 §2.3.3: §3.11 (heuristics).

P. 8 §2.3.4 2nd sentence: SB_3 miscite W_3. W_3 (§9 $\ell\ell$. 3-4 wrote "was common in preceding studies (DWZ; Weber & Kirsner 1997)." W_3 (§5, especially Footnote 11) had given further explanations to these references, and other references. SB_3 (§2.3.4 2nd sentence) cite W_3's sentence but omit the references ("examples") between brackets and then misleadingly claim "unsupported by examples".

P. 8 §2.3.4 penultimate para: "global linearity": how could anyone ever claim to infer this from a restricted domain as in these experiments?

Pp. 8-11 §3: §3.6 (statistics)

Pp. 11-12 §4: §3.9 on DWZ and §3.8 on linear utility

P. 11 ℓ.-10/-9: Marginal utilities do NOT cancel in SB's analysis (see my comment on SB_3 p. 5 ℓ.6).

P. 11-12, para there: to suggest novelty over DWZ, SB ignore the big overlaps and focus on small details of difference: that DWZ have no probabilities and no Method 2. Both details are to SB's disadvantage: ambiguity as examined by8 DWZ is more interesting and newer than risk as examined by SB, and SB's method 2 is only a roundabout way (involving also Method 1) to test additivity of decision weights, done more efficiently in the degeneracy tests of DWZ (§3.9 and Appendix C).

P. 12 §5: §3.12 (Method 2 understood); §3.7 (preceding falsifications of PT); §3.6 (statistics)

P. 12 Cm_1 last para "novel and appropriate way": SB's findings are neither of these.

Pp. 12-16 §6: §3.11 (heuristics)

P. 12 Cm_2 ℓ.-6 "gathering additional data": fixing mistake SBM1 while keeping mistakes SBM2-SBM7 is a waste of time.

P. 13 §6.1: §3.10 (no viable alternative)

Pp. 13-14 §6.2.1: §3.11 (heuristics)

P. 13 last para: see my comment on p. 3 (Claim #5).

P. 14 §6.2.2: §3.11 (heuristics)

P. 14 3rd para: subjects had to make $21 \times 38 + 7 \times 26 = 980$ choices, confirming W_0.

P. 14 Cm_1 ℓ.-2/-1: SB_3 again erroneously claim to identify the unidentifiable (§3.1).

P. 14 §6.2.3: §3.11 (heuristics)

P. 14 §6.2.3 ℓ.9: Ramsey (1931) explained that SB's claimed compromise does not exist (W_0 §4).

Pp. 14-15 §6.2.4: §3.11 (heuristics).

P. 15 2nd para refers to Fig. 2 on SB_3's p. 16: Panel A only at first requires explanation, as given to subjects by Diecidue et al. (2007). For repeated use throughout an experiment it is highly preferable to Panel B, contrary to SB_3's suggestions.

P. 15 §6.2.4, last para: Diecidue et al. (2007) debriefed pilot subjects.

P. 15 §6.2.5: $\mathcal{S}$3.13 (Method 2 understood)

Pp. 15-16 §6.2.6 & §6.2.7: $\mathcal{S}$3.11 (heuristics)

Pp. 16-18 §6.3: $\mathcal{S}$3.2 (stake-size)

P. 17 Cm_1 ℓ. -9/-8: SB_3 incorrectly claim that stakes are about expected values.

P. 17 §6.3.2 penultimate sentence: SB_3 do not give the relevant confidence intervals (W_3 Footnote 17).

P. 17 Footnote 40 on SB_2's space limitations: right/no claims do not take more space than wrong claims. The declared aim of only fixing SBM1 while leaving all other mistakes of $\mathcal{S}$2 is a useless exercise.

P. 18 Cm_2 top: whereas SB do not admit any mistake, I have to admit one[17]: W_3 (p.3 ℓ. -5) had forgotten that the unit of payment in Diecidue et al. (2007) was Dutch guilder, DFL, rather than euro. However, SB_3's calculations there are incorrect. Appendix C gives corrected calculations.

P. 18 Cm_2 3rd para. SB_3 write:

> "Wakker makes the patently false claim that "all statistical conclusions in BRS were based on accepted null hypotheses" (Section 9). On the contrary"

But Wakker's claim is correct, and nothing in SB_3's para or elsewhere contradicts it! SB_3 further write:

> "the figure [in SB_2] depicting our main results included confidence intervals, and the associated table included standard errors, making it easy for the reader to see that we obtained reasonably precise zeros" (p. 18)

I invite everyone to check out that it is not "easy to see". With only accepted H_0s[18], this point is crucial and should have been discussed. W_3 (Footnote 17) pointed out that SB_2 do not give the relevant confidence intervals. Importantly, W_3's §9 has further criticisms of SB_2.

P. 18 Cm_2 2nd & 3rd para (cherry picking): see *$\mathcal{S}$3.5*.

P. 18 last para: $\mathcal{S}$3.13 (Method 2 understood).

Pp. 19-21 §7: $\mathcal{S}$3.10 (no viable alternative)

P. 19 $\ell\ell$. 13-15 ("people prefer lotteries with fewer outcomes"): SBM4b and *$\mathcal{S}$3.3*.

Pp. 19-20 §7.1: $\mathcal{S}$3.4 cites Wakker (2023b), who carefully documented the correct definition of OPT (deviating from SB_3's unfounded claims).

P. 19 para -3 middle:

Analysis of Subtle Miscitation

SB_3 write on Wakker (2023b):

> by quoting a key passage from K&T, but in doing so changed singular nouns and verbs to plural. These changes *alter* the passage's meaning in a manner that *suggests* greater generality than the original text. [italics added]

By omitting relevant info (that Wakker explicitly indicated where he changed K&T's passage) and using two incorrect verbs ("alter" iso "generalize", and the implicit "suggests" instead of the explicit "provides") SB_3 mislead readers to think that Wakker misbehaved. Here is Wakker's (2023b p. 186) exact quotation of Kahneman and Tversky (1979):

---

[17] My second plea of guilt, and the only real one.

[18] Except a trivial monotonicity test of k_+ versus k_-.

> … prospects are segregated into two components: (i) the riskless component, i.e., the minimum gain or loss ... which is certain to be obtained or paid; (ii) the risky component, i.e., the additional gain[s] or loss[es] ... which is[are] actually at stake.... That is, the value of a strictly positive or strictly negative 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).

Wakker indeed *explicitly* indicated the changes made. The very only changes made were that, to go to multiple nonzero outcomes, some singulars *have to be changed* into plurals. These changes provided the only extension possible, leaving the meaning unaltered but, yes, with greater generality, as this is what generalizations do. W_3 handled the case fully correctly, did not "alter" or "suggest" anything, and SB_3 misled their readers to think otherwise.

P. 19 para -4: Contrary to SB_3's unfounded claim, Kahneman and Tversky did not "jettison" any generalization proposed in their 1975 paper, but called it "straightforward" (p. 288) (rather than problematic).

P. 19 ℓ. -4 on egregious implication: W_3 (§2.2) also indicated egregious implications.

P. 20 4th para: SB_3 again mention the mislabeling of SPT, again without admitting (§3.4).

P. 20 5th para: SB's opinions om the number of "formula(tion)s" that a functional consists of can only interest some non-mathematicians. See also SB_3 p. 19 para -3 and SB_0 Footnote 3.

P. 20-21 §7.2: §3.3 (complexity)

P. 21 3rd para: W_3 (§6) pointed out that what SB_3 call framing is a pure test of their complexity aversion. But SB_3 ignore the evidence (§3.3).

P. 21 4th para: W_3's OA.3 showed that SB_3's complexity aversion cannot eliminate any kind of SPT's anomalous violations of stochastic dominance. SB_3 counter that their complexity aversion can reduce the frequencies of such violations. I have seen stronger motivations for introducing a new theory!

P. 21 4th para: §3.3 explains W_3's term complexity aversion.

References

Abdellaoui, Mohammed, Chen Li, Peter P. Wakker, & George Wu (2020) "A Defense of Prospect Theory in Bernheim & Sprenger's Experiment", working paper; https://personal.eur.nl/wakker/pdf/abd.li.wak.wu_bernh.sp.pdf [W_0]

Aczél, János (1966) "*Lectures on Functional Equations and Their Applications.*" Academic Press, New York.

Bernheim, B. Douglas, Rebecca Royer, & Charles Sprenger (2022) "Robustness of Rank Independence in Risky Choice", *AEA Papers and Proceedings* 112, 415-420; https://doi.org/10.1257/pandp.20221090 [SB_2]

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; https://doi.org/10.3982/ECTA16646 [SB_0]

Bernheim, B. Douglas & Charles Sprenger (2023) "On the Empirical Validity of Cumulative Prospect Theory: A Response to the Wakker Commentaries", *Journal of Behavioral and Experimental Economics* 107, 102120; https://doi.org/10.1016/j.socec.2023.102120 [SB_3]

Birnbaum, Michael H. (2008) "New Paradoxes of Risky Decision Making", *Psychological Review* 115, 463–501; https://doi.org/10.1037/0033-295X.115.2.463

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; https://doi.org/10.1007/s11166-007-9011-z.

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. https://doi.org/10.1146/annurev-economics-080511-110950

Kahneman, Daniel & Amos Tversky (1975) "Value Theory: An Analysis of Choices under Risk", paper presented at the ISRACON conference on Public Economics, Jerusalem, 1975; http://personal.eur.nl/wakker/refs//pdf/ktpt75.pdf.

Kahneman, Daniel & Amos Tversky (1979) "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47, 263–291; https://doi.org/10.2307/1914185.

Quiggin, John (1982) "A Theory of Anticipated Utility", *Journal of Economic Behaviour and Organization* 3, 323–343; https://doi.org/10.1016/0167-2681(82)90008-7.

Quiggin, John & Peter P. Wakker (1994) "The Axiomatic Basis of Anticipated Utility; A Clarification", *Journal of Economic Theory* 64, 486–499; https://doi.org/10.1006/jeth.1994.1078

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.)

Schmeidler, David (1989) "Subjective Probability and Expected Utility without Additivity," *Econometrica* 57, 571–587;
https://doi.org/10.2307/1911053

Sonsino, Doron, Uri Benzion, & Galit Mador (2002) "The Complexity Effects on Choice with Uncertainty—Experimental Evidence", *Economic Journal* 112, 936–965;
https://doi.org/10.1111/1468-0297.00073

Tversky, Amos & Daniel Kahneman (1992) "Advances in Prospect Theory: Cumulative Representation of Uncertainty", *Journal of Risk and Uncertainty* 5, 297–323;
https://doi.org/10.1007/BF00122574

Wakker, Peter P. (2023a) "A Criticism of Bernheim & Sprenger's (2020) Tests of Rank Dependence", *Journal of Behavioral and Experimental Economics* 107, 101950; [W3]
https://doi.org/10.1016/j.socec.2022.101950

Wakker, Peter P. (2023b) "The Correct Formula of 1979 Prospect Theory for Multiple Outcomes", *Theory and Decision* 94, 183–187;
https://doi.org/10.1007/s11238-022-09885-w

Weber, Elke U. & Britt Kirsner (1997) "Reasons for Rank-Dependent Utility Evaluation", *Journal of Risk and Uncertainty* 14, 41–61;
https://doi.org/10.1023/A:1007769703493