Reply to Commentaries: Why should we worry about scientific conjunction fallacies?

Michel Regenwetter

Corresponding author. Departments of Psychology, Political Science, Electrical & Computer Engineering, University of Illinois at Urbana-Champaign

regenwet@illinois.edu. 603, E.Daniel Str., Champaign, IL61820

Maria M. Robinson

Department of Psychology, University of California San Diego

mrobinson@ucsd.edu

Cihang Wang

Department of Economics, University of Illinois at Urbana-Champaign

cwang153@illinois.edu

Reply to Commentaries:
Why should we worry about scientific conjunction fallacies?

Disclosures and Acknowledgments

All authors have read and approved the final manuscript. This work was supported financially by National Science Foundation grant SES # 20-49896 to Regenwetter (PI) and Army Research Office MURI Grant W911NF-20-1-0252 (PI: C. Langbort, Co-PIs: M. Başar, M. Regenwetter). ARO and NSF had no other role besides financial support. The authors are not aware of any conflicts of interest. The work in this paper has not previously been presented at any meetings.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARO, NSF, the U.S. Government, colleagues, or the authors' home institutions. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

This manuscript was written in LaTeX.

Regenwetter, Robinson, and Wang (in press a) is one-half of a larger project. A companion paper (Regenwetter, Robinson, & Wang, in press b) identifies four internal inconsistencies in Tversky and Kahneman (1992). It then lays out in broader terms how behavioral science tends to use the scope of theories (e.g., "critical tests") in a one-sided fashion to point out limitations of existing theory, but rarely delineates the intended scope of new theories. It explains how both the theoretical scope and the parsimony of theories like Cumulative Prospect Theory are inherently ambiguous. Because much of that companion paper provides an in-depth answer to various points raised in the Commentaries here, we do not dwell on those points in this response.

Response to Erev and Feigin (in press)

Erev and Feigin distinguish two methods for capturing heterogeneity: *individual-first* and *distribution-first* approaches. As we understand the former, the analyst estimates parameters separately for each individual in a study, then uses those inferences to predict a distribution of behaviors over individuals. In the second approach, as we understand it, the analyst estimates parameters of some population distribution, then uses those inferences to predict behaviors of individuals. How do these approaches connect to scientific logical reasoning errors like the conjunction fallacies that we warn about? We would argue that neither approach is immune to incorrect logic. In the first approach, the analyst should take precautions to avoid fallacies of composition in which they might draw incorrect inferences from the specific to the general. In the second approach, the analyst should guard against fallacies of sweeping generalization in which they might draw unwarranted inferences from the general to the specific. For an in-depth discussion of these two problems, see Regenwetter and Robinson (2017).

Erev and Feigin suggest that distribution-first approaches perform better in prediction tournaments. The authors do not provide enough details for us to take a specific stance on that conclusion. We would conjecture that the method used to assess "performance" in prediction tournaments can be biased towards favoring one or the other

approach. For instance, consider the number of "correct predictions," which is a common measure of performance. Also, for a concrete illustration, take Erev, Ert, Plonsky, Cohen and Cohen's (2017, p.4) car equipped with a novel device that enhances safe driving. Suppose we consider 100 'randomly selected' features (analogous to 'random stimuli' and even 'randomly selected behavioral phenomena') that are of varying importance to passenger safety. Suppose we install the device and we observe that it operates with near perfection on 98 of these 100 features. However, 0.5% of drivers will end up driving the wrong way on a freeway exit ramp at some point (based on the fail rate of feature #99). A different population of 0.5% of drivers will inadvertently drive onto crowded sidewalks (due to the fail rate of feature #100). We would have scores of dead with this system, even though nearly $\frac{9{,}999}{10{,}000}$ of its predictions are correct. Counting correct predictions compresses so much information across people and stimuli that it is susceptible to the fallacy of choice tallies (see also Regenwetter & Robinson, 2017, p.535). In fact, similar to the points we make in our paper (Regenwetter et al., in press a), a theory's predictions may be assessed very favorably via a "correct predictions" tally even when every individual violates that theory. Because the number of "correct predictions," tallied across individuals and stimuli is logically disconnected from an "individual-first" approach, that particular statistic, for instance, would lack a rationale as a performance measure for "individual-first" approaches in prediction tournaments.

Erev and Feigin also highlight that the "individual-first" approach has a number of parameters smaller than the number of independent observations per participant. As we spell out in the paper [add page numbers at time of publication], we would qualify this insight: 1. The number of parameters need not be smaller than the number of degrees of freedom in the data. 2. While this can affect identifiability of parameters, it need not make a model un-parsimonious. 3. In the context of theory testing, rather than behavior prediction, models can have vastly more free parameters than there are degrees of freedom in the data, yet still be extremely restrictive and readily falsifiable.

Response to Scheibehenne (in press)

Scheibehenne states that "deliberately casting individual differences as error variance and relying on median statistics and group-level aggregations yields more reliable predictions that are less prone to unsystematic noise." We disagree. Much of our paper precisely aims to dispel this broadly held view as a myth. It is important to note that our paper largely circumvents statistical inference, as we formulate our arguments at the level of (hypothetical) population distributions and population parameters. Consider the simple example of rolling a fair die, for which we know the 'population' distribution. The median and mean are 3.5. Casting the integers 1, 2, . . . , 6, as "errors" around that median or mean leads to reliably incorrect predictions: We actually expect to *never* observe a 3.5 roll, and predicting die rolls has nothing to do with "errors" of any kind. Reinterpreting genuine individual differences as "errors" or "noise" is akin to abstracting away from the faces of the die and therefore, potentially, abstracting away to something that is reliably incorrect, and always descriptive of nobody. Yet, to connect back to our response to Erev and Feigin, it could appear to perform well when assessed empirically through an aggregate lens. To be very clear: It is not an educated guess that a randomly sampled driver from a randomly selected country drives roughly in the middle of the road, give or take some noise (as though the side a country drives on were an error). Nor is it an educated guess that a random location in the United States experiences average annual rain fall (typically, it is either long term drought or frequent flooding). Representativeness heuristics can be dangerous misconceptions. In the case of behavioral decision research, we contend that they are deeply engrained and extremely difficult to root out.

We also disagree with Scheibehenne's characterization that "Regenwetter et al.'s paper addresses . . . essentially a trade-off between generalizability and goodness-of-fit" and the claim that "accounting for all individual differences risks overfitting." The latter statement contradicts the technical definition of overfitting, which, simply put, is fitting a model to error variance (noise) in data, rather than capturing substantive variation.

Consider the fair die again. The non-integer median and mean are both complete misrepresentations of any possible outcomes. Characterizing a die as having six equally likely faces that show the integers 1, 2, . . . , 6 does not, at all, trade off between generalizability and goodness-of-fit. Because the six sides of the die do not take the form of noise or error around the median/mean, accounting for all six outcomes has nothing to do with overfitting. As we move from a six-sided die to a parametrized decision theory or a collection of behavioral phenomena, conceptualizing heterogeneity as a theoretical primitive, rather than an ad-hoc add-on noise term, as a matter of theory building, is also a priori orthogonal to issues of overfitting noise in empirical data. To be clear: Predicting that drivers (in various countries) will drive either on the left or the right side of the road is not overfitting anything, nor does predicting that opposite coasts of the United States will continue to suffer from different types of extreme weather have anything to do with overfitting. For a related discussion, see Hertwig and Pleskac's (2018) concerns about a bias-variance dilemma and Regenwetter and Robinson's (2019) reply to those concerns.

A particularly noteworthy part of Scheibehenne's comment are his figures, which particularly nicely illustrate our core message: There is a huge diversity of probability weighting curves inferred from the data. Almost none closely align with the stylized curve associated with CPT_{MED}.

Scheibehenne advocates that "the ultimate measure against overgeneralizations and conjunction fallacies lies in better theorizing and the development of causal models of behavior." We fully agree with the first recommendation: Indeed, we advocate that heterogeneity of behavior should be treated as a theoretical primitive rather than an ad-hoc add-on to decision theory. The second recommendation seems to us conjecture. In Cumulative Prospect Theory, utility functions and probability weighting functions are hypothetical constructs that are not directly observable and that must be inferred from data. Causal and cognitive models are quite similar. They share the same property: Attention, memory, horse races, cognitive resources etc., are also hypothetical constructs

that we cannot observe directly. Just like inferences about utilities and weights can be subject to conjunction fallacies, so are cognitive constructs susceptible to Linda problems. It is a popular belief among scholars that cognitive constructs benefit from high face validity and can be revealed through "process" measures. This ideal must be weighed against the fact that cognitive models often bridge the construct-behavior gap with many, and very strong, auxiliary assumptions (e.g., relating reaction times to processing speed, or gaze to attention). These technical assumptions are arguably not an integral part of substantive psychological theory. Occasional violations of such assumptions may wreak havoc on replicability (Regenwetter & Cavagnaro, 2019). The most severe Linda problems come about when inference about processes relies heavily on (often parametric) data aggregation, because the latter could arguably offer more opportunities for aggregation artifacts to take hold and for such artifacts to be perpetuated through replication. We do not mean to question cognitive, causal, or process models wholesale. Rather, we merely wish to clarify that our paper applies to cognitive research paradigms just the same as it does to behaviorist models.

Response to Kellen (in press)

In his commentary, Kellen asks us not to use the term "Linda problem" for the types of scientific conjunctions we discuss. We grant Kellen that we are not aware of any decision scholars who explicitly state that decision makers are more likely to satisfy CPT_{MED} than they are to satisfy CPT. Hence, we are not aware of papers that overtly assign the event "the person satisfies CPT" a smaller probability than a strictly nested sub-event "the person satisfies CPT_{MED}." However, Linda problems are not so limited. Suppose that participants merely indicate their preferred characterization among "Linda is a bank teller" or "Linda is a feminist bank teller," rather than being asked to state explicitly which is *more likely*. To us, this is still a Linda problem, even though it does not offer participants a chance[1] to commit an explicit verbatum violation of the rules of

[1] Nor does it force on them what seems like a trick choice between violating probability theory or

probability theory. Likewise, advocating that 'people' satisfy CPT_{MED} is, to us, still a conjunction fallacy like those that occur in Linda problems. As a matter of fact, such a claim also still strains probability theory because it effectively advocates that 'people' satisfy a point hypothesis which, through a Bayesian lens, has probability zero, and which, through a frequentist lens, is refutable at any level of statistical significance if we just gather enough data. We all know that modes can be very different from both means and medians: Our best guess at the roll of a die is not a 3.5, because that is literally an impossible outcome. We show why our best guess at the behavior of a person should not be CPT_{MED}. It is a big stretch to suggest that many, most, or all individuals satisfy a conjunction of stylized properties such as those embodied in CPT_{MED}. In sum, we stand by our assessment that behavioral decision research is a Linda problem because behavioral decision research needs to navigate the conjunction or co-occurrence of behavioral phenomena. When communicating behavioral decision insights to policy makers and managers, we should take every precaution to guard against either committing, communicating, or encouraging conjunction fallacies.

Second, Kellen criticizes how we use prior data for illustration when making our conceptual points. When considering choice proportions from Birnbaum (2008), Erev et al. (2017), Kahneman and Tversky (1979) and Tversky and Kahneman (1992), we implicitly use numerous simplifying assumptions. For instance, we treat everyone's study samples as representative of the population, participants as properly motivated, stimuli as diagnostic and informative, study measures as valid and reliable, and we treat the observed proportions of phenomena as proxies of population probabilities. We make these simplifying assumptions deliberately, for the sake of a conceptual analysis, and to give prior scholars the benefit of the doubt in numerous regards. That is, in contrast to Kellen's assessment to the contrary, we would consider our assumptions to be charitable views of prior work. Granted, one could maintain that *every individual* satisfies the conjunction of

---

discarding valuable information.

all phenomena in *each* paper by allowing sufficiently unreliable measures, sufficiently bad quality participants, sufficiently undiagnostic stimuli, and allowing any number of other imperfections in the data. But we would not consider that a "charitable" view of prior studies. Kellen draws attention to one of these numerous simplifications, the fact that we treat observed proportions as proxies of population proportions. He then asserts that we "assume that erroneous responses are *impossible*." It seems to us self-evident that Kellen considers non-representative samples, poorly motivated study participants, non-diagnostic stimuli, unreliable measures, etc. well within the realm of possibility, even if he does not discuss those. Likewise, we believe it should be clear that we make our assumptions for illustrative purposes. Kellen proceeds to consider a study in which every individual, for every stimulus, has an identical probability of 0.3 of committing an error in their response. He then recalculates that rather than 57% of people showing the combination of "new choice paradoxes" in Birnbaum (2008), that number would, instead, be 68%. The calculation is correct[2], but the argument is one-sided. To give an example, if, instead, we were to assume that respondents committed errors with probability 0.7, then the proportion of people showing the conjunction of phenomena would drop to about one-third. In short, our examples are conceptual. Taking into account response errors and other aspects of a study, such as the representativeness of the sample, reliability of measures, diagnosticity of stimuli, etc., can shift these numbers around in any direction, in ways big or small. We prefer not to distract the reader with speculations on how these problems factor in, or how they trade off with each other. We omit these topics in our illustrative examples for the simple purpose that we wish to stay on message: Stylized characterizations of behavior, especially conjunctions of stylized phenomena, are subject to conjunction fallacies. Behavioral decision researchers should take care not to commit, communicate, or encourage conjunction fallacies when teaching others about how 'people' make decisions.

Third, Kellen suggests that more people show stylized phenomena than can be

---

[2] We thank David Kellen for locating a calculation error in our manuscript.

revealed in a given study because there likely exist some stimuli, not used in a given study, in which those phenomena will reveal themselves on individuals who have not shown the phenomena already. That is, the original stimuli may not be fully diagnostic. This is a very good point. Upon close inspection, Kellen's argument adds additional weight to our warnings: If a theory's stylized predictions are contingent on idiosyncratic circumstances because the stimulus must be 'just right' for a person to show a given phenomenon, then we should warn members of the public of this caveat when we tell them how 'people' make decisions. In particular, citing Kellen's insight, we should warn policy makers that one needs 'just the right combination of just the right circumstances' to experience the conjunction of stylized properties that much behavioral decision research routinely advertises wholesale.

Kellen also reminds us of the useful role of "recipes" for designing new studies that can place formidable pressure on old theories. Our companion paper (Regenwetter et al., in press b) critiques this wide-spread practice in great detail. Here, we merely consider the connection between "recipes" and scientific conjunction fallacies. Our point in this paper (Regenwetter et al., in press a) is that scholars routinely and strongly overstate the co-occurrences of phenomena by prominently relying on conjunctions: Suppose that a recipe for smart experimentation enables us to detect that 60% of individuals have property X (far from everybody) and another recipe reveals that 60% have property Y (far from everybody). But suppose that only 10% combine both properties. If scholars commit the conjunction fallacy of suggesting that "most" people have properties X and Y, then this is both wrong and strongly misleading. In such a case, the recipes obstruct our very ability to see that almost everybody is an exception to the X-AND-Y pattern. Matters become much worse with collections of more than two phenomena. Recipes are not in any way immune to conjunction fallacies, nor do they protect against them. For instance, what would be a recipe to make 'people' become intransitive? Take choice options that have three competing attributes, find one person who cares only about Attribute 1, according to

which A is the best, B middle, and C is the worst among three options. Find one person who cares only about Attribute 2, according to which B is the best, C is middle, and A worst. Find one more person who cares only about Attribute 3, according to which C is the best and B is the worst. Weak stochastic transitivity is violated! A randomly selected person among the three most likely prefers A to B. A (separately, with replacement) randomly selected person among the three most likely prefers B to C. Yet, a (separately, with replacement) randomly selected person among the three most likely prefers C to A. On the surface, it looks like we might have found a recipe for revealing intransitive preference. But, because we baked in a conjunction fallacy by design, our recipe is rather a recipe for creating artifacts (in this case a Condorcet paradox). Each of the three decision makers has a transitive preference (notably ABC, BCA, or CAB). Granted, Kahneman and Tversky (and others) provided recipes for revealing "pervasive effects." This really still changes nothing about conjunction fallacies regarding the joint occurrence of such "pervasive effects." In other words, "recipes" for revealing "pervasive effects," per se, offer no protection against scientific conjunction fallacies.

Kellen is correct to remind us that "critical tests" challenge a "received view." Decades after the publication of Kahneman and Tversky (1979) and Tversky and Kahneman (1992), there can be little doubt that CPT_{MED} and the conjunction of stylized properties have become the received view of many. Our paper is a 'critical test' of the logic behind such a received view. Kellen states that "stylized characteristics serve to show that some people's choices deviate from the received view, to a degree that cannot be ignored." We are likewise argue that the striking difference between stylized characterizations (today's received view) and genuine heterogeneity should also not be ignored. Indeed, Kellen suggests that, perhaps, the wide-spread use of CPT_{MED} as a way to represent Prospect Theory is merely "a habit or a desire to be consistent with past treatments." We agree that this may well be the case, and this was indeed a major motivation to take stock of that received view. We aim not to speculate about scholar's motivations, nor do we

allege bad intentions. Rather, we raise concerns about the cost to society of (intentionally or not) characterizing human decision making in overly stylized and misleading ways.

Kellen concludes that "the complexity of the subject matter calls for careful historical, methodological, rhetorical, and conceptual considerations that go beyond what RRW currently have to offer." Indeed, we could not agree more strongly. While our paper, together with its companion paper, attends to some historic components in the discussion of (C)PT and some other prominent papers, while we discuss some methodological and conceptual matters, much work remains to be done. We are particularly pleased that like-minded colleagues such as Erev, Feigin, Kellen, and Scheibehenne are interested, willing, and highly qualified to help lead behavioral decision research beyond stylized theory and away from logical reasoning fallacies.

References

Birnbaum, M. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*, 463–501.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*, 369-409.

Erev, I., & Feigin, P. (in press). Heterogeneous heterogeneity: Comment on Regenwetter, Robinson, and Wang (in press). *Decision*.

Hertwig, R., & Pleskac, T. (2018). The construct–behavior gap and the description–experience gap: Comment on Regenwetter and Robinson (2017). *Psychological Review*, *125*, 844-849.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

Kellen, D. (in press). Behavioral decision research is not a Linda problem: Comment on Regenwetter, Robinson, and Wang (in press). *Decision*.

Regenwetter, M., & Cavagnaro, D. (2019). Tutorial on removing the shackles of regression analysis: How to stay true to your theory of binary response probabilities. *Psychological Methods*, *24*, 135-152.

Regenwetter, M., & Robinson, M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*, 533-550.

Regenwetter, M., & Robinson, M. (2019). The construct-behavior gap revisited: Reply to Hertwig and Pleskac (2018). *Psychological Review*, *126*, 451-454.

Regenwetter, M., Robinson, M., & Wang, C. (in press a). Are you an exception to your favorite decision theory? Behavioral decision research is a Linda problem! *Decision*.

Regenwetter, M., Robinson, M., & Wang, C. (in press b). Four internal inconsistencies in Tversky and Kahneman's (1992) Cumulative Prospect Theory paper: A case study in

ambiguous theoretical scope. *Advances in Methods and Practices in Psychological Science*.

Scheibehenne, B. (in press). Experimenter meets correlator: Comment on Regenwetter, Robinson, and Wang (in press). *Decision*.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323.