Four internal inconsistencies in Tversky and Kahneman's (1992) Cumulative Prospect Theory paper: A case study in ambiguous theoretical scope and ambiguous parsimony

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, IL 61820

Tel.: 1 217 3330763, Fax: 1 217 244 5876, Web: www.regenwetterlab.org

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

Abstract

Scholars heavily rely on theoretical scope as a tool to challenge existing theory. We advocate that scientific discovery could be accelerated if far more effort were invested into also overtly specifying and painstakingly delineating the intended purview of any proposed new theory at the time of its inception. As a case study, we consider Tversky and Kahneman (1992). They motivated their Nobel Prize winning *Cumulative Prospect Theory* with evidence that, in each of two studies, roughly half of the participants violated *independence*, a property required by *Expected Utility Theory.* Yet, even at the time of inception, new theories may reveal signs of their own limited scope. For example, we show that Tversky and Kahneman's findings in their own (1992) test of *loss aversion* provide evidence that at least half of their participants violated their theory, in turn, in that study. We highlight a combination of conflicting findings, in the original paper, that make it ambiguous to evaluate both Cumulative Prospect Theory's scope and its parsimony, on the authors' own evidence. The Tversky and Kahneman (1992) article is illustrative of a social and behavioral research culture in which theoretical scope plays an extremely asymmetric role: To call existing theory into question and motivate surrogate proposals.

Keywords: Cumulative Prospect Theory, Median Responses, Theoretical Scope, Theoretical Parsimony

Four internal inconsistencies in Tversky and Kahneman's (1992) Cumulative Prospect Theory paper: A case study in ambiguous theoretical scope and ambiguous parsimony

Disclosures and Acknowledgments

This work was supported financially by National Science Foundation grants SES-14-59699 and SES # 20-49896 to Regenwetter (PI) and the Army Research Office under MURI Grant Number W911NF-20-1-0252 (PI: C. Langbort, Co-PIs: M. Başar, M. Regenwetter). NSF and ARO had no other role besides financial support.

M. Regenwetter guided and supervised the project and contributed most of the writing. M. Robinson and C. Wang discovered the inconsistencies when exploring how much preference heterogeneity Cumulative Prospect Theory permits. They provided most of the mathematical proofs in joint work while M. Robinson was a PhD student in Psychology at Illinois and C. Wang an undergraduate Mathematics/ Economics double major at Illinois. All authors have read and approved the final manuscript.

The authors are not aware of any conflicts of interest.

Part of this paper has previously been presented orally at the annual Society for Mathematical Psychology satellite meeting of Psychonomics in 2019, and at the 57th Annual Edwards Bayesian Research Conference in 2020.

Thanks to Meichai Chen, Brittney Currie, Yu Huang, Emily Line, Xiaozhi Yang, and to our referees for comments on earlier drafts. We are grateful to Lyle Regenwetter for providing an elegant mathematical proof of Claim I. 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 the Army Research Office or the U.S. Government, the National Science Foundation, of colleagues, or of 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.

Introduction

As Paul Meehl (1978) famously stated, theories in many “areas of Psychology lack the cumulative character of scientific knowledge. They tend neither to be refuted nor corroborated, . . . ” (abstract). While Meehl was primarily referring to paradigms that grow into and out of fashion, we contend that the statement applies to some of the most famous and enduring scholarly contributions. Also, while Meehl was primarily referring to what he called “soft” areas of psychology, his characterization can apply to high-profile and mathematically formal theories. In this article, we consider one of the most prolifically cited papers in behavioral science, as a case study in both ambiguous theoretical scope and ambiguous parsimony. That paper uses theoretical scope as a tool to challenge prior theories, yet a closer look at the paper’s own evidence calls the paper’s own scope into question by the very same criteria. Both proponents and opponents of this theory can cherry-pick the ways in which they characterize its scope and/or its parsimony, based solely on the original paper, in ways that serve their goals. In some domains of science, scholars have reached broad consensus about many theories’ “edge conditions,” their actual scope, as well as their flexibility. Behavioral research is yet to find even minimal consensus about whom its theories describe, and under what circumstances.

The paper is organized as follows. We start by examining what it might mean to specify theoretical scope and parsimony, first with a simple example from the natural sciences and then with a few exemplary papers from Psychology. This prepares the ground for our case study, the 1992 paper by Daniel Kahneman and Amos Tversky on Cumulative

Prospect Theory. In the context of that paper, we discuss the difference between theoretical scope and parsimony in more detail. In particular, we lay out how scholars can cherry-pick aspects of the same evidence on the same theory to draw diametrically opposite conclusions. Next, we dig more deeply into Tversky and Kahneman (1992) to discuss four internal inconsistencies within that single seminal paper. Citing the example of such a stellar theory, we make our case that social scientists need to consider the intended scope and the parsimony of their theoretical proposals much more carefully and with less bias. We then turn to the general question of how behavioral scientists can clarify the scope and parsimony of their theories. We review symptoms that flag problems, and we sketch some ideas for better practice.

What is Theoretical Scope? What is Parsimony?

Physicists share virtually perfect consensus about the *theoretical scope* of Newton's Law of Gravity: The law applies to all objects in all locations at all times, as long as the objects in question are not "too small" and do not "move too fast." Much of Engineering is based on understanding the situations in which Newton's Laws of Motion, Pascal's Law of Pressure, Boyle's Law of Gases, etc. either do or do not apply verbatum. There is also broad consensus on how one can cover a broader range of phenomena using more complex (i.e., less *parsimonious*) theories.[1]

What is theoretical scope and parsimony in Psychology? If we consider any

[1] For an interesting case study drawing a parallelism between Newton's laws and a well known cumulative program of research in psychology, see Navarro (2021), and an earlier paper by Shepard (1987), on which it builds.

behavioral regularity, whether it is in cognition, personality, social interaction, or another domain, is there a consensus in the field as to who displays this behavior and under what circumstances? Do the Behavioral Sciences strive to develop broad agreement on delineating the range of conditions in which a theory applies, and the characteristics of people whose behavior it explains?

Imagine that Archimedes' Principle was subject to major exceptions: If the weight of the displaced water only matched some boats' weight, how would that affect boat design? Psychologists take it for granted that their theories only hold with exceptions. The logic of permissible exceptions is baked into much of our statistical methodology: Whenever a statistically significant proportion of participants in a study, but nowhere close to all, show a certain behavioral regularity, say, they remember more, or take larger risks, or are more cooperative, we infer that the phenomenon or effect is 'real.' But what does that tell us about who actually satisfies that regularity when, how, and why? If it is too ambitious to characterize people to this level of detail, do Psychologists at least estimate how large a proportion of the population obeys their hypotheses about decision making, memory, perception, personality, reasoning, or social interaction? How does the discipline identify and interpret conditions, individuals, stimuli, or tasks, where a theoretical claim does not actually apply? In other words, how does Psychology conceptualize theoretical scope and how does it handle limitations in a theory's scope? Going one step further, how does the field ensure the parsimony of new theories that aim to encompass phenomena not captured by earlier theories?

To illustrate current best practice in stating the intended purview of new theories, we briefly consider three very recent high-profile papers from cognitive psychology, namely Popov and Reder (2020) in long-term memory research, Schneegans, Taylor, and Bays (2020) in working memory research, and Lleras et al. (2020) in visual attention. These are three research paradigms in which individual differences may plausibly play a muted role, compared to, say, clinical, developmental or social psychology. Popov and Reder (2020) proposed a new theory and computational model that purports to explain a wide range of "frequency effects" on memory, while also providing a process-level explanation for how working memory capacity gives rise to these effects. Popov and Reder (2020) acknowledged (p. 38): *" . . . our extensions to recall tasks could be considered lacking in some respects. For example, while our serial recall model does a good job capturing the interaction between word-frequency and serial position, it does not reproduce the one item recency effect, which has been attributed to access from a WM buffer (Anderson, et al., 1998). Furthermore, we have made no attempt to model the full specter of contiguity effects in free recall, which have been a crucial benchmark for models of free recall . . . "* Schneegans et al. (2020) proposed a framework grounded on a specific neural model of visual working memory. A purported strength of this model is that it links visual working memory limits to a concrete neural substrate. Schneegans et al. (2020) discussed limitations of their theory (pp. 8-9), stating, e.g.: *"In keeping with most previous work on VWM limits, we have not here attempted to reproduce the variations in bias and precision that are observed for different feature values . . . "* Lleras et al. (2020) proposed a novel visual search model, in part to account for the specific form of "search functions." They reported that their theory could account for a wide range of effects in visual search tasks, particularly the effects of the similarity between target and distractor items in the visual search array for simple and real-world objects. Lleras et al. (2020) acknowledged (p. 422): *"A . . . limitation is that TCS is currently mostly focused on parallel processing and efficient search. More work is needed to flesh out what happens after parallel evidence accumulation is stopped and several target likely locations need to be inspected."*

These papers stand out in that, unlike many scholarly papers in Psychology, they at least acknowledge phenomena that their theories do not explain. However, while they do take the admirable step of reviewing limitations, these papers are nonetheless reflective of a research culture where theoretical scope is usually treated like a set of moving goal posts. Oftentimes, stating that "more work is needed" can be a diplomatic way to acknowledge that the scope of a theory is inherently ambiguous. For instance, even in these cognitive tasks, the extent to which every effect holds in every individual person and across a broad range of contexts is ambiguous. Hence, it is not really clear how desirable it is to model all effects jointly in the same individuals, without committing a conjunction fallacy, for instance. Even in these exemplary papers, the challenging question of how to weigh their own theory's limitations against its ability to account for an enlarged range of phenomena remains unanswered. Schneegans et al. (2020) used some heuristic statistical measures of parsimony (known as AIC and BIC) based on counting the number of free parameters in the theory. The other two papers compared their own work with other work conceptually.

Ultimately, despite their best efforts, all three papers are ambiguous about both the scope and the parsimony of their theories. Our article aims to raise awareness of the asymmetric role that theoretical scope plays in the development of social and behavioral theory.

The Asymmetric Role of Theoretical Scope

Notwithstanding the higher level of nuance in the above three papers, it is common to use theoretical scope almost exclusively to motivate new theories.

Oftentimes, the scope of an existing theory is delineated through “critical tests,” or “unaccounted-for effects,” by selecting an empirical paradigm and carefully designing certain stimuli, for which a substantial number of people generate data that conflict with the existing theory’s predictions. By focussing on critical tests or unaccounted-for effects, scholars deliberately place pressure on existing theory. Proposing a new theory that passes those same hurdles creates an inherent bias in favor of the new theory. By the design of the research paradigm, this bias is immune to detection by even the best standard statistical model selection criteria. This is because model selection methods typically apply post-hoc, only after the scholar has already selected a suitable paradigm and crafted the relevant diagnostic stimuli to stress-test the old theory. Yet, support for a new theory may, in fact, already be ambiguous at the time of its inception, because some participants may already provide some evidence against the new theory on some of the stimuli. Indeed, since presumably no behavioral theory performs universally well for everyone on all stimuli and in all contexts, a new difficulty arises as soon as the new theory reveals some of its weaknesses. Now, different schools of thought may disagree on whether the cracks in the new theory represent new critical tests that create the need for yet another theory, or whether these fissures are merely examples of the imperfections that are inherent to even the best behavioral theories.

In this paper, we unpack these ideas for one paper that proposed the most prominent theory of decision making: Cumulative Prospect Theory. Even though the authors did not highlight them, the first fissures are already visible in the original paper proposing the theory. As for how to weigh this theory's improvement over prior theories against its own limitations, that question, even decades later, has neither been settled nor has it been discussed in much depth.

Cumulative Prospect Theory: A Case Study

Few theories from the Social and Behavioral Sciences are as prominent across all of science, and even popular science, as Cumulative Prospect Theory (CPT). Yet, since its inception, CPT has also become a routine lightning rod for countless competing proposals about decision behavior. The theory enjoys extremely broad use in applied settings, where it guides much policy development, while also being the target of abundant skepticism, especially in basic research. While some critics call it extremely narrow and easy to refute, others think it is too flexible and even irrefutable. We consider both scenarios later in this paper. One possible explanation for finding both strong support and strongly mutually contradictory criticisms of one and the same theory might be that this theory may perform extremely well in accounting for some people's behavior in some circumstances, but not others. Intuitively speaking, the theory may have limited scope.

Tversky and Kahneman (1992) premised Cumulative Prospect Theory on showing that many people display phenomena in violation of prior models such as Expected Utility Theory (see also Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). Specifically, Tversky and Kahneman (1992) reported two studies in which 53% (money managers) and 46% (Stanford students) of the participants violated a core property of Expected Utility Theory called *independence* (pp. 303-304, and their Tables 1 and 2). Let E denote the event that the Dow Jones changes by a certain amount between today and tomorrow. Suppose that two lotteries f and g both yield $25,000 if E occurs. Suppose that f' and g' are the same as f and g, respectively, except that they both yield $0 if E occurs. If event E occurs, it does not matter whether the decision maker has chosen f or g today. Indeed, if E occurs, then they receive $25,000 tomorrow. Likewise, if E occurs, then it does not matter whether they have chosen f' or g' today, since f' and g' cause them neither to win nor to lose anything tomorrow. Through the lens of Expected Utility Theory the outcomes under event E simply 'cancel out.' Therefore, anyone who prefers f to g must prefer f' to g' and vice-versa, because these preferences should only depend on what happens when something other than E occurs, and in that situation, f and f' are identical to g and g', respectively. However, as we already mentioned, 53% of Tversky and Kahneman's money managers chose f over g and chose g' over f'. If roughly half of all participants violate a property required by standard theory, Tversky and Kahneman argued, then a new theory is needed. Later in this paper, we show that Tversky and Kahneman's own (1992) findings elsewhere in the same paper provide evidence that, in turn, at least half of all participants in their test of another property, called *loss aversion*, likewise violated their own theory. This and other incongruences in Tversky and Kahneman (1992) raise questions about how we should conceptualize CPT's theoretical scope.

The goal of reporting these disparities is not to take a stance about either the validity or value of CPT, or the lack thereof (recall that we focus on one paper, Tversky & Kahneman, 1992, not an entire research program). In our view, every behavioral theory has limited scope. Rather, we aim to highlight the one-sidedness in much behavioral science, where theoretical scope is disproportionately used to censure a leading theory, often with little discussion of the intended or expected scope of the new theory. Building on earlier preparatory work (Davis-Stober & Regenwetter, 2019; Regenwetter & Robinson, 2017), we advocate that Social and Behavioral Science should develop more constructive ways to reconcile, synergize, and weigh the theoretical proposals associated with different schools of thought. These recommendations reach beyond the replication crisis to more general questions, including the need to think through unintended consequences of successful replication (see also Davis-Stober & Regenwetter, 2019; Irvine, 2021; Kellen, 2020; Regenwetter & Robinson, 2017; Rotello, Heit, & Dubé, 2015; Yarkoni, 2020).

What is the Difference between Theoretical Scope and Parsimony?

Consider a prospect that leads to outcome A with probability p and outcome B otherwise, where A and B are monetary amounts. Positive values for A and/or B are monetary gains, whereas negative values are monetary losses. Table 1 shows eight such lotteries, taken from Tversky and Kahneman (1992). For example, the first column shows a

Our label	I	II	III	IV	V	VI	VII	VIII		I	II	III	IV	V	VI	VII	VIII
	Eight Loss Prospects:									Eight Loss Prospects:							
Outcome A	-50	-100	-200	-200	-400	-100	-150	-200		-50	-100	-200	-200	-400	-100	-150	-200
p	0.1	0.05	0.01	0.1	0.01	0.1	0.05	0.05		0.1	0.05	0.01	0.1	0.01	0.1	0.05	0.05
Outcome B	0	0	0	0	0	-50	-50	-100		0	0	0	0	0	-50	-50	-100
$1-p$	0.9	0.95	0.99	0.9	0.99	0.9	0.95	0.95		0.9	0.95	0.99	0.9	0.99	0.9	0.95	0.95
EV	-5	-5	-2	-20	-4	-55	-55	-105		-5	-5	-2	-20	-4	-55	-55	-105
	Predicted preferences according to CPT:									Predicted preferences according to the “toy” theory:							
	Equations 1 and 3								Row $\sum$	Eqs. 1 and 3 with the modification that $1 < \beta, \delta < 2$							
	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	8	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	0	7								
	1	1	1	1	1	1	0	0	6								
	1	1	1	1	1	0	1	0	6								
	1	1	1	1	1	0	0	0	5								
	1	1	0	1	0	1	1	1	6								
	1	1	0	1	0	1	1	0	5								
	1	1	0	1	0	1	0	0	4								
	1	1	0	1	0	0	0	0	3								
	1	0	0	1	0	1	0	0	3								
	1	0	0	1	0	0	0	0	2								
									6	0	1	1	0	1	1	1	1
									6	0	0	1	0	1	1	1	1
									3	0	0	0	0	0	1	1	1
									2	0	0	0	0	0	1	0	1
									1	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1

Preference patterns among Loss gambles of Tversky and Kahneman's Table 3 prospects with $p \leq 0.1$. A "1" indicates "risk-seeking" preference in favor of a lottery over the sure amount equal to the lottery's expected value (EV). The left side shows the predictions under Eqs. 1 and 3, whereas the right side shows predictions according to Eqs. 1 and 3 according to a "toy" theory where $1 < \beta, \delta < 2$. The predicted number of "risk-seeking" choices is tallied in the center column. The preference pattern predicted by Expected Utility Theory is underlined.

lottery in which the decision maker runs a 10% chance ($p = 0.1$) of losing \$50 (i.e., $A = -50$), otherwise (with probability $1 - p = 0.9$) neither winning nor losing anything. This lottery has an expected value (EV) of $0.1 \times (\$-50) = (\$-5)$, as mentioned in the table. In line with Tversky and Kahneman's terminology, choosing to play this lottery over paying \$5 (its EV) is called a *risk-seeking* choice. The opposite choice is called *risk-averse.*

Expected Utility Theory (EUT) transforms dollar amounts to subjective utilities and calculates the expected value of the subjective utilities to determine whether it is preferable to play this lottery or (in this example) to accept the sure loss. Using a utility function for losses of the form $v(x) = -(-x)^{\beta}$, where $0 \leq \beta < 1$, EUT implies that, in each lottery of Table 1 the lottery is preferable to the sure loss of paying the lottery's EV (regardless of the value of the parameter $0 < \beta < 1$). Accordingly, showing a risk-seeking preference as a “1,” the first preference pattern in the table contains a string of ones across the eight lotteries. This preference pattern is underlined in the table. Disregarding the knife-edge possibility of being indifferent between a lottery and its EV, there are $2^8 = 256$ possible binary preference patterns for eight pairwise decisions. Of these, EUT with the above utility function only permits one preference pattern here: At least as far as these lotteries are concerned, EUT (with that utility function) is rather parsimonious. If everyone consistently acted in a risk-seeking fashion when confronted with any of these choices, then EUT would also have excellent scope, again, at least as far as these lotteries are concerned. However, if there are one or more lotteries among the eight where some, many, or all people make risk-averse choices (i.e., they choose to pay the sure loss of the EV, rather than play the lottery) then EUT suffers from limited scope: In that case, either EUT may only apply at best to certain decisions but not others, or it may only capture the behavior at best of some people but not others, etc.

Having established limitations in scope of a given theory, it is common (while not ubiquitous) that scholars will propose a less parsimonious revision that includes the original theory as a special case. Such is the case with CPT: It contains EUT as a special case, but CPT introduces extra flexibility that accommodates a broader range of possible preference patterns. As stated in the 1992 paper, CPT transforms money into subjective utilities in a fashion similar to the above version of EUT, it transforms probabilities into subjective weights, and it switches to a cumulative weighted average calculation. We omit the details of the latter as they are not important here. To keep mathematical formulae to a minimum we only restate CPT's core building blocks: subjective utility and probability weighting. According to Eqs. 5-6 of Tversky and Kahneman (1992), CPT invokes a value function v and probability weighting functions w^+ and w^- for gains and losses, with parameters $\alpha, \beta, \gamma, \delta, \lambda$, of the form

$$v(x) = \begin{cases} x^\alpha & \text{if } x \geq 0 \qquad (0 \leq \alpha \leq 1), \\ -\lambda(-x)^\beta & \text{if } x < 0 \qquad (0 \leq \beta \leq 1; 0 < \lambda), \end{cases} \tag{1}$$

$$w^+(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}} \qquad (0 < \gamma \leq 1), \tag{2}$$

$$w^-(p) = \frac{p^\delta}{(p^\delta + (1-p)^\delta)^{\frac{1}{\delta}}} \qquad (0 < \delta \leq 1), \tag{3}$$

where x are money amounts and p are probabilities. See Tversky and Kahneman (1992) for the details on how they combined these functions to derive subjective utilities of lotteries and to model preferences among lotteries.

For now, we only consider the left half of Table 1. It shows twelve different preference patterns predicted by CPT, for those eight prospects. Since there are no mixed prospects (i.e., lotteries involving both gains and losses), we set $\lambda = 1$ without loss of generality, thereby effectively dropping it from Equation 1 and reducing the utility function to the same form we used for EUT. Setting $\beta = 0.44$ and $\delta = 0.9$, for example, generates the first pattern (also generated by EUT), setting $\beta = 0.35$ and $\delta = 0.89$ generates the second pattern in the table. We derived the 12 preference patterns by plugging 1,000 distinct values greater than zero and smaller than one, for each of β and δ (1 million combinations of values) into Equations 1 and 3. Rather than the single pattern predicted by Expected Utility Theory, CPT accommodates 12 out of the 256 possible binary preference patterns for these eight decision problems[2]. In other words, as far as these decision problems are concerned, while arguably still parsimonious, CPT is less parsimonious than EUT. If more people made decisions in accordance with these 12 preference patterns rather than just the one predicted by EUT, then, at the price of a reduction in parsimony, CPT would have purchased greater theoretical scope. If, similarly to various laws of Physics or Chemistry, 'everyone in a certain well-delineated population' made decisions in accordance with these 12 preference patterns, and, more generally, made decisions consistent with CPT for 'all decisions of a certain well-delineated type,' then we would have a clear sense of CPT's (possibly 'immense') theoretical scope. On the flip side, the cost in parsimony might not be commensurate with the enhancement of its theoretical scope if 'too many' people violated CPT in 'too many situations.'

One thing is clear: Permitting more than one preference pattern is an important step towards enhancing theoretical scope. Tversky and Kahneman (1992) reported extensive individual differences in behavior, a finding that virtually all scholars agree with, at least in principle. Put differently, the theoretical scope of any theory of decision making hinges, at least in part, on its ability to balance individual differences with theoretical parsimony. The combination of a single set of mathematical formulae (Equations 1-3) with parameters that can each vary across a continuum of permitted values, creates the

[2] We assume, for simplicity, that our grid search found all preferences patterns that are possible.

potential for a theory that is both parsimonious and enjoys great scope[3].

Tversky and Kahneman (1992) summarized some of their key findings as follows (pp. 311-312):

> *The median exponent of the value function was 0.88 for both gains and losses, in accord with diminishing sensitivity. The median λ was 2.25, indicating pronounced loss aversion, and the median values of γ and δ, respectively, were 0.61 and 0.69 [. . .] The parameters estimated from the median data were essentially the same.*

Following Regenwetter and Robinson (2017), we refer to CPT with these parameter values as CPT_{MED}. Returning to Table 1, a decision maker who satisfies CPT_{MED} will prefer to incur the sure loss rather than play the lottery in every one of the eight decision problems. That predicted preference pattern, shown at the bottom of the table, is the diametrical opposite of the EUT prediction. Together, these 12 preference patterns provide an example of CPT's ability to model huge diversity in a parsimonious fashion, including two preference patterns that are diametrical opposites. It may be tempting to cite CPT's ability to produce two diametrically opposite preference patterns as 'proof' that it lacks parsimony, and maybe even that it 'can fit anything.' However, since the table still rules out 244 out of 256 binary preference patterns, such speculation would be premature.

We have already asserted that scholars often use theoretical scope in a one-sided

[3] Replacing $\alpha, \beta, \gamma, \delta, \lambda$ by jointly distributed random variables models within- and/or between individual variation, as well as all preferences permitted by Equations 1-3 (see, e.g., Kellen, Pachur, & Hertwig, 2016; Murphy & ten Brincke, 2018; Nilsson, Rieskamp, & Wagenmakers, 2011; Pachur, Schulte-Mecklenberg, Murphy, & Hertwig, 2018; Regenwetter et al., 2014; Scheibehenne & Pachur, 2015; Zwilling et al., 2019).

fashion to establish limitations in the scope of someone else's theory. We have also mentioned that CPT has served as a lightning rod for alternative proposals about decision making. In fact, it is common practice to reduce CPT down to CPT_{MED} and to question the entire theory by offering evidence that many people violate the predictions of CPT_{MED} on some stimuli (see, e.g., Birnbaum, 2008; Brandstätter, Gigerenzer, & Hertwig, 2006, for particularly prominent examples of this line of reasoning in support of alternate proposals)[4]. To see how counter-productive this approach can be, consider Table 1 once more. Because EUT is a special case of CPT, anyone who satisfies EUT also satisfies CPT. Yet, anyone who satisfies EUT has the exact opposite preference from CPT_{MED} for every one of the lotteries in the table. Hence, an EUT decision maker (who, by design, also satisfies CPT) violates *every* prediction of CPT_{MED} in the table. Using the number of violations to CPT_{MED} as a measure of performance of CPT as a whole would lead one to conclude falsely that '*every* prediction of CPT is violated,' when considering the preference pattern of an EUT decision maker. While unilaterally challenging the theoretical scope of one theory to advance another theory already stacks the odds against existing theory, the common practice of attacking CPT by testing only CPT_{MED} goes much further in that it misrepresents CPT while calling its scope into question.

We round out our comparison of theoretical scope and parsimony by showing that defining the latter is just as elusive as defining the former. Consider Table 1 once more.

[4] Note that, besides CPT_{MED}, Brandstätter et al. (2006) also considered two other versions of CPT "with prior parameters." Their "Lopes & Oden (1999)" version, where $\beta = 0.0.97, \delta = 0.99$ yields the single pattern 11111000 in the fifth row of our Table 1. Their "Erev et al. (2002)" version provided values only for α and γ. For the corresponding gain gambles, out of the corresponding 12 mirror image preference patterns, both their "Lopes & Oden (1999)" ($\alpha = 0.55, \gamma = 0.70$) version and their "Erev et al. (2002)" ($\alpha = 0.33, \gamma = 0.75$) version of CPT give the single pattern 00000111.

We derived preference patterns also for a hypothetical “toy” theory that satisfies Equations 1-3, but with parameter ranges $1 < \alpha, \beta, \gamma, \delta < 2$. The resulting preference patterns for the same Loss gambles are given in the right half of Table 1. In one sense, this theory contradicts CPT: Diametrically contrary to CPT, it uses convex utility for gains and concave utility for losses. Yet, both CPT and the “toy” theory use the same mathematical formula and both also use four parameters that are each defined on a unit interval. Hence, by the most popular heuristic conceptualization of model complexity, both theories are equally parsimonious. Yet, for these stimuli, the “toy” theory only predicts 7 binary preference patterns rather than CPT’s 12, two of which are shared with CPT, namely the pattern also predicted by EUT, and the pattern also predicted by CPT_{MED}. In the next section, we show that matters are even more ambiguous and that one could make a case, using this empirical paradigm and these stimuli, that CPT is more, equally, or less parsimonious than the “toy” theory.

Four inconsistencies

We now discuss four internal inconsistencies within Tversky and Kahneman (1992) and how they impact our understanding of CPT’s theoretical scope. As these inconsistencies reveal, Tversky and Kahneman provided evidence that their own theory suffers from limited scope almost in the same way as EUT. More generally, the inconsistencies reinforce the point that social scientists need to meticulously consider the intended scope of their own theory and how evidence stacks for or against it, rather than disproportionately focus on the limited scope of contending theories.

Inconsistency (i):

Tversky and Kahneman (1992, p.306) stated:

> “The most distinctive implication of prospect theory is the fourfold pattern of risk attitudes. For the nonmixed prospects used in the present study, the shapes of the value and the weighting functions imply risk-averse and risk-seeking preferences, respectively, for gains and for losses of moderate or high probability. Furthermore, the shape of the weighting functions favors risk-seeking for small probabilities of gains and risk aversion for small probabilities of loss, provided the outcomes are not extreme.”

Similar statements appeared in the paper’s abstract. The paper aimed to document this fourfold pattern in their Table 4, using four different types of prospects described in their Table 3. We provide an adapted excerpt of that table in our Table 2. All of these prospects were of the form $(A, p; B, 1-p)$. Our Table 1 shows the collection of eight Loss lotteries with $p \leq 0.1$ that underlie the data in Column 4 of Tversky and Kahneman’s Table 4. The corresponding eight Gain lotteries with $p \leq 0.1$ underlying their Column 2 of their Table 4 can be obtained by taking the absolute values of all the outcomes in our Table 1. Tversky and Kahneman also utilized a collection of 17 Gain lotteries and a collection of 17 Loss lotteries, all with $p \geq 0.5$, for Columns 3 and 5 of their Table 4. We do not repeat those stimuli here.

According to the fourfold pattern, for these prospects, decision makers make risk-seeking choices for Loss prospects with $p \leq 0.1$ and Gain prospects with $p \geq 0.5$, whereas they make risk-averse choices for Gain prospects with $p \leq 0.1$ and Loss prospects with $p \geq 0.5$. Tversky and Kahneman's Table 4 reports, separately for each of 25 subjects, the percentages of risk-seeking choices in Gains with $p \leq 0.1$ in Column 2, Gains with $p \geq 0.5$ in Column 3, Losses with $p \leq 0.1$ in Column 4, Losses with $p \geq 0.5$ in Column 5 (as well as aggregated results). Again, see our Table 2 for an adapted excerpt. Perfectly error-free adherence to the fourfold pattern would mean entries of 100 in Columns 2 and 5, and 0 in Columns 3 and 4.

Column 1	Column 2	Column 3	Column 4	Column 5
Subject	Gain lotteries	Gain lotteries	Loss lotteries	Loss lotteries
	$p \leq 0.1$	$p \geq 0.5$	$p \leq 0.1$	$p \geq 0.5$
4	71%	0%	30%	58%
6	100%	5%	0%	100%
21	100%	0%	0%	100%
Average	78%	10%	20%	87%

Table 2
Percentages of risk-seeking choices among Subjects 4, 6, 21, and average percentage, adapted from Table 4 of Tversky and Kahneman (1992).

As we saw earlier, we derived the 12 preference patterns in Table 1 by plugging 1,000 distinct values greater than zero and smaller than one, for each of β and δ into Equations 1 and 3. Plugging the same 1 million values for α and γ into Equations 1-2 generates twelve preference patterns for the eight Gain lotteries with small p by switching 0's and 1's in the lower part of Table 1. The center column of our Table 1 shows the associated number of risk-seeking choices, predicted in an error-free decision maker, for each preference pattern. For any integer N from 0 to 8 (except 1) we have found values of β, δ with which CPT predicts a percentage of $\frac{N \times 100}{8}$ in Column 4 of Tversky and

Kahneman's Table 4 (and in our Table 2). Replacing losses by gains in the top and switching around 0's and 1's in the lower part of our table, shows that for any number N' from 0 to 8 (except 1) we have found values of α, γ for which CPT predicts $\frac{(8-N')\times 100}{8}$ in their Column 2. Taken together, CPT permits almost[5] any possible combination of values in Columns 2 and 4 of their Table 4, even in error-free choice.

This finding leads to several noteworthy insights: 1. Because Equations 1-3 can accommodate almost any conceivable data in Columns 2 and 4 of their Table 4, CPT does not actually imply a fourfold pattern on the prospects that Tversky and Kahneman used in their fourfold pattern study. In particular, the theory is less parsimonious than advertised. 2. Any scholar who only focussed on these stimuli and on the number of predicted risk seeking choices among them, might mistakenly infer that CPT is a nearly vacuous and essentially irrefutable theory. 3. On these stimuli, the number of risk seeking choices is a metric that obstructs a clear view of both CPT's scope and its parsimony. 4. The stimuli we reproduced in Table 1 (and which took from the original paper) are not diagnostic of CPT's empirical performance, when viewed through the lens of the number of risk seeking choices.

As we show in the Proofs section, Equations 1-3 do, however, indeed predict 0 in Column 3 and 100 (%) in Column 5 of Tversky and Kahneman's Table 4 (our Table 2), almost regardless[6] of the parameter values in Equations 1-3. This finding also leads to

[5] Note that, while we did not find cases where $N = 1$ or $N' = 1$, this might be possible with parameter values not included in our grid search.

[6] We need to avoid $\alpha = \gamma = 1$, and avoid $\beta = \delta = 1$, which is the reduction of CPT to Expected Value theory. We also need to avoid $\alpha = 0$ and avoid $\beta = 0$, the cases of constant utility.

several noteworthy insights: 1. On these stimuli, a portion of the fourfold pattern does indeed follow from the theory. In particular, with respect to the stimuli in Tversky and Kahneman's Columns 3 and 5, the theory is, indeed, extremely parsimonious. 2. Any scholar who only focussed on these stimuli and on the number of predicted risk seeking choices among them, might mistakenly infer that CPT is extremely narrow and easy to refute. 3. On these stimuli, the metric of counting risk seeking pairwise preferences happens to be informative because a value of 0 (or 100) implies a risk averse (or seeking) preference for every stimulus. 4. These stimuli are extremely diagnostic of CPT's empirical performance, as assessed through the number of risk seeking choices. In particular, just six out of 25 participants (including Subject 21 in our Table 2) were perfectly aligned with CPT's prediction in that they showed 0 in Column 3 and 100 in Column 5. To label the other 19 participants as "consistent" with CPT, one needs to permit anywhere from 5% (Subject 6, Column 3) to 42% (Subject 4, Column 5) response errors. 5. One way to evaluate CPT's scope on these stimuli, would be to develop a model of within- and between-person variability that allows us to infer who, or how many people, satisfy/violate the theory's predictions on which stimuli, after accounting, e.g., for response errors.

In all, Tversky and Kahneman's discussion of fourfold patterns creates an internal tension in that, 1) in contrast to their claim, their theory does not actually imply a fourfold pattern on the stimuli they used to document that pattern; 2) their theory is virtually immune to rejection on two columns in their Table 4 because of undiagnostic stimuli (or an undiagnostic performance statistic); and at the opposite extreme, 3) the other two columns can only be viewed as supporting CPT if one is willing to permit substantial error rates in responses in many participants. In all, Tversky and Kahneman's own study of fourfold patterns paints an ambiguous picture about both the parsimony and the theoretical scope of their theory.

This leads us back to considering the inherent difficulty, in current day Psychology, to properly define what we mean by either *parsimony* or *scope* of a theory and how we can go about assessing either of these concepts. One way of looking at this challenge is to consider the substantial ambiguity associated with asking what constitutes a suitable study design to either reject or support a theory: Tversky and Kahneman used two studies, each with just two decision problems, but a large number of participants, to challenge Expected Utility Theory's scope in their test of independence. In contrast, their assessment of CPT involved far fewer participants but many more stimuli, for each study. How do these experimental design choices affect the balance of evidence between competing theories? A contemporary statistical analysis could evaluate the parsimony and statistical power of each approach, given a suitable probabilistic specification of the theory, for given stimuli. It is far less clear how it would evaluate and take into account the very design of the tasks, studies, and stimuli themselves (see also Broomell & Bhatia, 2014, for important related challenges). The perplexing tradeoff between 'diagnostic' and 'undiagnostic' stimuli is particularly striking in the fourfold pattern study.

In this context, it is useful to return to the "toy" theory we have discussed earlier. We mentioned earlier that both theories use the same mathematical fomulae, the same number of parameters, and two different but equally sized (unit interval) domains for their parameters. Yet, we also saw that CPT permits almost twice as many binary preference patterns on these stimuli. We omit a proof, but, when concentrating only on the number of risk seeking choices, the "toy" theory can accommodate almost any imaginable data in Tversky and Kahneman's Table 4 with little or no reliance on response errors. This makes the "toy" theory almost irrefutable by that particular statistic on those stimuli. In contrast, as we have seen, through the lens of the same statistic, CPT is more parsimonious in that it predicts values of 0 in Column 3 and 100 in Column 5. This creates a tension: CPT can be viewed as equally parsimonious, more parsimonious, or less parsimonious than the "toy" theory, depending on the view point. At the same time, by almost any measure of fit, CPT does not fit the data nearly as well as the "toy" theory, in two columns. Comparing both the scope and the parsimony of CPT versus the "toy" theory would only become more complicated if we were to expand from these stimuli to include other stimuli, from binary choices and lotteries to include other tasks, or from these participants in this lab to include other participants in other labs. All in all, we face an inherent ambiguity as to which theory performs better and at what cost. Psychology is yet to develop agreed-upon ways to weigh experimental task, study design, stimulus design, parsimony, and goodness-of-fit. Many aspects of this tradeoff reach beyond model selection methods in Statistics, largely because the concept of degrees of freedom is ill-defined without a specified data structure. For related points, see Yarkoni (2020) who warned that sampling individuals from a population, sampling stimuli from a universe of possible stimuli, and sampling tasks or other study features from a design space, creates many additional, and unaccounted for, sources of variance. "Failing to model such factors appropriately (or at all) means that a researcher will end up either (a) running studies with substantially higher-than-nominal false positive rates, or (b) drawing inferences that technically apply only to very narrow, and usually uninteresting, slices of the universe the researcher claims to be interested in." (Yarkoni, 2020, p. 5)

Returning to Tversky and Kahneman (1992), besides the fourfold pattern, they also discussed a phenomenon called *loss aversion*, according to which decision makers may go to great lengths to avoid losing something. Tversky and Kahneman asked decision makers to determine the value of x that makes a 50/50 chance of receiving either \$$a$ or \$$b$ equally attractive as a 50/50 chance of receiving either \$$c$ or \$$x$. Our Table 3 shows the eight decision problems they used (see also their Table 6). In many cases, but not all, some among a, b, c are negative numbers, hence denoting monetary losses. The three remaining inconsistencies, which we discuss next, are all related to that study. Each, again, highlights the asymmetric and ambiguous role of theoretical scope in Tversky and Kahneman's paper.

Inconsistency (ii):

We first re-evaluate Tversky and Kahneman's findings for their Problem 7 (where $a = 50, b = 120, c = 20$), for which they reported a median x value of 149. Suppose for a moment that they correctly identified the median of population x values with perfect accuracy. A median x of 149 would mean that, given the choice between a 50/50 chance of winning either \$50 or \$120, on the one hand, and a 50/50 chance of winning either \$20 or

\$149, half of the population with either be indifferent or prefer (\$50, $\frac{1}{2}$; \$120, $\frac{1}{2}$), and half of the population would either be indifferent or prefer (\$20, $\frac{1}{2}$; \$149, $\frac{1}{2}$).

Alas, it is impossible, according to CPT, to be indifferent between (\$50, $\frac{1}{2}$; \$120, $\frac{1}{2}$) and (\$20, $\frac{1}{2}$; \$x, $\frac{1}{2}$), for any $x \leq 149$. As we show in the Appendix, it is mathematically impossible to obtain a value of $x \leq 149$ for Problem 7 using Equations 1-3 above[7]. This means that the theory is incompatible with half of the population having x values smaller than 149, as would be implied by the definition of a median. In other words, if we take Tversky and Kahneman's median x of 149 in Problem 7 at face-value, then, just as roughly half the participants violated prior theories in Tversky and Kahneman's test of independence, so did at least half the participants violate CPT in their test of loss aversion. For this particular decision problem, CPT suffers from almost exactly the same limitation in theoretical scope as the limitation that Tversky and Kahneman (1992) used to hamstring Expected Utility Theory (using two decision problems). Yet, this limitation in CPT's theoretical scope, to our knowledge has not been previously discussed in the literature. Inconsistency (ii) illustrates the strikingly asymmetric role of theoretical scope in decision research: It is routine to cite Tversky and Kahneman's findings only to question EUT but not to question CPT.

[7] Note that, say, allowing $\alpha, \beta > 1$, as we entertained in the "toy" theory, say, for half the population, as a way to resolve this inconsistency, would conflict with the reported median parameter estimates, according to which at least half the population has $\alpha, \beta \leq 0.88$, thereby creating a different incongruency.

Inconsistency (iii):

We now consider Problem 8, where $a = 100, b = 300$, $c = 25$, and for which Tversky and Kahneman (1992) reported a median x of 401. Here, we find a different type of internal inconsistency. Tversky and Kahneman, and many scholars since, like to characterize the theory and empirical findings through summary statistics, such as median parameter estimates and median responses. Their median x of 401 in Problem 8 does not directly contradict CPT as a whole. However, when taken at face value, it is nonetheless incompatible with their reported median γ value of 0.61, regardless of the other parameter values in CPT. We provide a demonstration in the Appendix.

Problem	a	b	c	Median x	Median θ	x from CPT_{MED}	θ from CPT_{MED}
1	0	0	-25	61	2.44	68.5	2.74
2	0	0	-50	101	2.02	137	2.74
3	0	0	-100	202	2.02	274	2.74
4	0	0	-150	280	1.87	411	2.74
5	-20	50	-50	112	2.07	132	2.73
6	-50	150	-125	301	2.01	357	2.76
7	50	120	20	149	0.97	169	1.63
8	100	300	25	401	1.35	429	1.72

Table 3
Test of loss aversion in Tversky and Kahneman (1992). For each problem, participants chose a value of x that made the prospect $(a, .5; b, .5)$ equally attractive as $(c, .5; x, .5)$. The fixed values of a, b and c, median values of x, median values of $\theta = |(x - b)/(c - a)|$ are from Tversky and Kahneman (1992). The values of x and θ derived from CPT_{MED} are shown in the right two columns.

Inconsistency (iv):

Table 3 summarizes some information of Tversky and Kahneman's direct test of loss aversion (see their Table 6). Besides providing median x values they also reported that the median value of $\theta = |(x - b)/(c - a)|$ approximately equaled 2 in Problems 1-6. In their view, in addition to a median λ of 2.25, this empirically supports loss aversion and the popular stylized claim that 'losses loom roughly twice as large as gains.' We challenge these pervasive interpretations and conclusions using Tversky and Kahneman's (1992) own published data.

As we quoted earlier, Tversky and Kahneman reported (p. 312), without giving any details, that the "parameters estimated from the median data were essentially the same" as CPT_{MED}. We derived x and θ values from CPT_{MED} (right two columns of Table 3). The monetary amounts are systematically larger than the median x values, especially strongly so in Problem 4. Likewise, switching from x to θ, the values of θ predicted by CPT_{MED} are systematically larger than their empirical counterparts. In all, Tversky and Kahneman's (1992) median responses, median parameter estimates, and theory are internally misaligned. To our knowledge, it is an open question how to combine these different points of view on Tversky and Kahneman's own empirical evidence into a concise assessment of the theory's scope. Inconsistency (iv) tells us that even very broad statements, e.g., about the size of the subpopulation whose parameters satisfy various stylized properties about risk attitudes and/or probability weighting, are not well founded in the original CPT paper. In particular, it is not at all clear, from their study, for whom, for how many people, and for which decisions, it is actually the case that "losses loom twice as large as gains." In our view, this inherent ambiguity of theoretical scope has huge policy implications. It is not at all clear, from Tversky and Kahneman's or other scholars' evidence, who is served, and who is hurt, by policies built on the presumption that losses loom about twice as large as gains.

How can we clarify scope and parsimony?

Recently, there has been much activity to protect Psychology from fraud, improve the quality of research, and strengthen theory. This work has led to prominent recommendations for good practice with respect to a variety of goals. We now review and comment on some of these from our perspective of theoretical scope and parsimony.

One prominent recommendation is to preregister studies. Preregistration is often promoted as a way to decrease post-hoc analyses and theorizing because it forces researchers to identify key hypotheses prior to data collection (e.g. Mistler, 2012; Moore, 2016; Simmons, Nelson, & Simonsohn, 2021; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). However, as noted by others, preregistration is not a panacea for poor theory development, mediocre methods, or undiagnostic data (see, e.g. Lakens & DeBruine, 2021; Szollosi & Donkin, 2021; Szollosi et al., 2020). It is unclear how preregistration guards against the problems and errors we have discussed here. Preregistered studies can engage in the same logical fallacies, can use the same stylized statistics, can perpetuate the same double standards, can repeat the same asymmetric philosophies of science, and can be as internally inconsistent as non-preregistered studies. Preregistering a design that perpetuates ambiguous scope and ambiguous parsimony merely documents study flaws in advance. On the up-side, preregistration provides an opportunity for scholars to discuss issues of scope, parsimony, diagnosticity of stimuli, fairness of model selection, etc., ahead of running a study, if they so choose.

A second prominent recommendation is increased emphasis on replication (see, e.g., Pashler & Harris, 2012; Simons, 2014). Replication helps to improve measurement precision and asses the reliability of an effect of interest. This is inherently useful and can also be leveraged to compute lower and upper bounds on the number of people who satisfy a theoretical claim or display a phenomenon (see, e.g., Bogdan, Cervantes, & Regenwetter, n.d.; Davis-Stober & Regenwetter, 2019; Heck, 2021). As such, replication can help assess scope. However, along with others we advocate that replicability is far from a panacea: For one thing, efforts invested into reproducing and replicating a prior study as identically as possible are efforts not invested into exploring how the finding extends to other people, novel stimuli, different tasks, or new contexts. Relatedly, for arguments on the relative merits, e.g., of direct and conceptual replication, see also Carpenter (2012); Nosek, Spies, and Motyl (2012); Pashler and Harris (2012); Schmidt (2009); Simons (2014). In other words, replication can be orthogonal to explorations of theoretical scope. Just as importantly, like preregistration, successful replications of a phenomenon would not guard against most of the errors we identify here, such as fallacies of sweeping generalization, conjunction fallacies, and other problems associated with stylized statistics. To the contrary, it can repeat, reinforce, and even perpetuate reasoning errors and scientific biases (Davis-Stober & Regenwetter, 2019; Irvine, 2021; Regenwetter & Robinson, 2017, 2019a, 2019b; Rotello et al., 2015; Yarkoni, 2020). On the up-side, we can envision situations in which scholars could both reproduce a prior study and enhance it with additional features that aim to bring theoretical scope and parsimony into better focus. We also advocate that scholars preface replication with a discussion of its impact on understanding scope.

A third recommendation, which is gaining traction especially in cognitive psychology, is to replace or supplement verbal theories with formal computational or mathematical models (e.g., Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021; Grahek, Schaller, & Tackett, 2021; Guest & Martin, 2021; Navarro, 2021; Oberauer & Lewandowsky, 2019; Robinaugh, Haslbeck, Oisín, Fried, & Waldorp, 2021; van Rooij & Baggio, 2020). We agree that formal modeling can force researchers to think more explicitly about both the intended scope and the flexibility of their theories. However, formal modeling on its own is not sufficient for addressing many of the double standards and ambiguity problems that we have identified. Clearly, CPT is a formal model. Yet, our discussion above demonstrates that the value of formal modeling hinges on how it is implemented (and this point is reinforced by all of the above references). Formal modeling can give the appearance of rigor and mask systemic errors (Chen, Regenwetter, & Davis-Stober, 2021). Formally precise models oftentimes force simplifying assumptions or omit hidden variables. These can become counter-productive (see, also Kellen, Davis-Stober, Dunn, & Kalish, 2021; Yarkoni, 2020, for related general points). To keep formals models tractable, scholars may limit themselves to overly simple tasks or simple stimuli (see, e.g., Navarro, 2021, for a discussion). On the up-side, in contrast to verbal theories, which we would consider inherently ambiguous, formal modeling does provide a common language (logic, computer code, mathematics, and/or statistics) through which to discuss theoretical scope, parsimony, and standards of scientific discourse openly and rigorously. However, as we see when we review the fifth recommendation, while mathematical formulae ostensibly eschew rhetoric, the connection between the mathematics and the substantive questions of interest can also be ambiguous. Our discussion of CPT in this paper highlights an example of that broad problem.

A fourth prominent recommendation is to supplement or replace data fitting with prediction to other tasks or unseen data (e.g., Busemeyer & Wang, 2000; Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Erev et al., 2010; Pitt, Kim, & Myung, 2003; Yarkoni & Westfall, 2017). This has been advocated as a tool for addressing overfitting and for developing theories that generalize. We agree that prediction is an important step towards recognizing and avoiding heuristic approaches to parsimony. However, to avoid asymmetries and double standards, scholars should provide a clear explanation why the participants, tasks, stimuli, and contexts are designed in such a way as not to provide an unfair advantage to some theories over others. Notice that searching a parameter space for best fitting parameters need not make a theory unparsimonious, nor need it cause overfitting. The number of parameters is no more than a heuristic measure parsimony[8]. CPT is a prominent example, where prediction has gone awry: 'Refuting' CPT by testing predictions from CPT_{MED} or other stylized distortions of the theory gives no consideration to either the theory's scope or its parsimony. In some prediction tournaments (e.g. Erev et

[8] Our example with EUT above made a deterministic prediction on those stimuli, despite having a 'free' parameter β. For an example of an extremely restrictive theory with 540 free parameters and 20 degrees of freedom in the data, see Regenwetter and Davis-Stober (2012). Viewed from a Bayesian perspective, the upper bound on the Bayes factor of this model against an unconstrained model is about two-thousand and the lower bound is zero. In the next subsection, our Inequalities 4-5 and equality constraints characterize a model with 144 free parameters, 50 degrees of freedom in the data, and an extremely parsimonious 12-dimensional prediction space of measure zero in the 50-dimensional empirical sample space.

al., 2010), even though the parameters of some theories could characterize specific properties of individuals[9], the tournament rules required participants to reduce their theory down to a single set of specific parameter values, thereby obfuscating individual differences, and collapsing the scope of each theory to a single set of stylized predictions. Similarly and more generally, claiming that a theory performs poorly in predictions is counter-productive when those predictions hinge on untested and unquestioned auxiliary assumptions, such as off-the-shelf statistical models. We turn to this next.

A final major recommendation is increased attention to the problem of *coordination* in psychological research: Theory simultaneously presumes and guides measurement of latent constructs (Irvine, 2021; Kellen et al., 2021; Singmann et al., 2021; van Frassen, 2008). Some of these papers warn that heated debates about the relative merit of competing theories often heed no attention to the pivotal role of technical and auxiliary assumptions, such as analysis-of-variance or other off-the-shelf models. Attending to the circular connection between theory and measurement (e.g., attending to auxiliary assumptions) forces researchers to consider both jointly. For a related literature, see the extensive work on meaningfulness in psychological measurement and theory (Falmagne & Doble, 2016; Falmagne & Narens, 1983; Narens, 2002, 2007; Roberts, 1985, 1998; Roberts & Rosenbaum, 1986). Attention to the coordination problem and to meaningfulness may

[9] Consider, for instance, the "Explorative sampler" model of Erev, Ert, and Yechiam (2008) in Erev et al. (2010). Equation 5 of Erev et al. (2010) contains a parameter δ describing a decision maker's sensitivity to the length of an experiment. Another parameter, k, captures the maximum number of past experiences a person samples from memory. The model also uses, e.g., a weighting parameter w and a "diminishing sensitivity" parameter β. While it would make sense that different people might have different values for these parameters, the tournament required a single value of each parameter to predict aggregated choice proportions. The performance measure aggregated and collapsed empirical information across decision problems, trials, and participants.

lead to a more nuanced understanding of both theoretical scope and theoretical parsimony.

Flagging symptoms of ambiguous scope or parsimony

We have touched on a number of features of Tversky and Kahneman (1992) whose parallels and analogues in other paradigms can flag ambiguous scope or ambiguous parsimony in psychological theory more broadly. 1) The most prominent flags are all forms of asymmetric reasoning, in which scholars point out shortcomings of others' theories or evidence without discussing the possible shortcomings of the replacements they propose. Double standards, such as using many more or many fewer stimuli to test the old theory than the new one, using stimuli (even if picked 'randomly') that pressure the old theory but not the new one, pushing a novel theory merely on the basis of its ability to accommodate some 'anomalies' that the old theory does not explain, all of these may create systematic biases against existing theory and in favor of the proposed new theory. When the latter is custom-designed to handle certain phenomena, it is important to also understand the associated cost in parsimony. Extreme forms of asymmetric reasoning occur when scholars provide evidence only against special cases of a theory (e.g., CPT_{MED}), thereby literally misrepresenting the theory they question. 2) Serious questions of scope arise with mathematical errors or omissions. For example, the mathematical model in Tversky and Kahneman (1992) does not actually imply a fourfold pattern on their own fourfold pattern study stimuli. 3) More broadly, any internal inconsistencies in reasoning can flag problems with parsimony and/or scope. A very troubling, yet extremely common practice, is strawman Null hypothesis testing. Testing hypotheses whose violation is a foregone conclusion (e.g., perfectly calibrated coin flipping as a Null model of behavior, the Null that two groups are identical) cannot legitimately provide evidence in favor of a proposed theoretical claim (see also Cohen, 1994; Meehl, 1978). More broadly, meaningless statistics, such as the 'number of correct predictions,' or the number of correct modal choices in decision making, generate useless evidence. 4) Claims of "converging evidence" are often unsubstantiated. Here, it is useful to see whether it is possible to calculate or estimate how many people satisfy the conjunction of evidentiary phenomena (see also Davis-Stober & Regenwetter, 2019; Regenwetter, Robinson, & Wang, in press). 5) Nontechnical readers should be aware that some commonly used model selection criteria, such as AIC and BIC, are heuristic in nature, and as such, may give an analysis a sheen of rigor that need not be warranted. A major improvement is the use of Bayes Factors, especially in cases where the researchers provide information on the possible range of Bayes factors for a given study.

How parsimonious is CPT?

We end with an illustration of Bayes factors as a quantitative measure of parsimony for CPT. We concentrate on the $8 + 8 + 17 + 17 = 50$ stimuli from Tversky and Kahneman's (1992) fourfold pattern study used in their Table 4 (we show and label some of them in our Table 2). As we have already seen, there are 12 possible preference patterns for the 25 Gains Prospects, and 12 possible preference patterns for the 25 Loss Prospects. We briefly consider two *probabilistic specifications* of CPT (with Equations 1-3) from Regenwetter et al. (2014) and Zwilling et al. (2019). According to the *aggregation-based* model, each individual has one of the $12 \times 12 = 144$ allowed patterns (out of $2^{50} > 10^{15}$

possible ones) as her single "true" preference state. If she prefers prospect f to prospect g, and if we allow her to make a response error with probability[10] at most τ, then she will choose f with probability $\geq 1 - \tau$. According to the *random preference* model, the probability of choosing f over g is the total probability of those preference patterns in which f is preferred to g, in an unspecified probability distribution over all 144 possible preference patterns. Considering just the prospects and preference patterns in our Table 1 it is noteworthy that a decision maker who prefers the lottery in Prospect I also does so in Prospect IV and vice-versa. The same holds for Prospects III and V. Moreover, in any pattern where the lottery is preferable in Prospect II, the lottery is also preferable in Prospect IV, but the converse does not hold. Using similar reasoning, or using polyhedral combinatorics[11], one can show that, no matter what probability distribution we consider over these preferences, writing P_i for the probability of choosing the lottery in Prospect i,

$$1 \geq P_I = P_{IV} \geq P_{II} \geq P_V = P_{III} \geq 0; \qquad P_{II} \geq P_{VII} \geq P_{VIII} \tag{4}$$

$$P_{IV} \geq P_{VI} \geq P_{VIII} \geq 0; \qquad P_V + P_{VI} \geq P_{VII}. \tag{5}$$

The remaining choice probabilities for the other 17 Loss Prospects are one. Similar constraints hold among the choice probabilities for the 25 Gain Prospects.

Regenwetter et al. (2018) and Zwilling et al. (2019) review how to calculate the range of possible Bayes factors between a given such model and an unconstrained

[10] When $\tau = \frac{1}{2}$ the decision maker is at least as likely to choose the preferred prospect as the alternative.

[11] We used Porta, available at http://comopt.ifi.uni-heidelberg.de/software/PORTA/

"encompassing" model. The aggregation-based model can generate a Bayes factor anywhere between 0 and $\frac{1}{144\times\tau^{50}}$. For $\tau = 0.5$, the upper bound exceeds 10^{12}, and for $\tau = \frac{1}{4}$, it exceeds 10^{27}. Because the random preference model predicts deterministic choice of the lottery in Tversky and Kahneman's (1992) seventeen "high-probability" Loss Prospects, as well as deterministic choice of the sure gain in Tversky and Kahneman's (1992) seventeen "high-probability" Gain Propects, the Bayes factor, either in favor, or against the random preference model, is unbounded! To conclude about CPT's parsimony through the lens of Bayes factors: For both of these probabilistic specifications, there is no limit as to how much evidence could be provided against CPT, on these stimuli.

Conclusion and Discussion

Since the publication of Prospect Theory some 40 years ago (Kahneman & Tversky, 1979), scholars have prolifically cited Tversky and Kahneman's (and others') findings that Expected Utility Theory suffers from limited scope. Yet, for nearly 30 years, it has gone unnoticed that Tversky and Kahneman (1992) provided evidence for the exact same limitation in CPT's scope: Just like half their participants violated Expected Utility Theory early in their paper, so did ostensibly half the respondents violate CPT in Problem 7 of their loss aversion study later in the same paper. In addition, some aggregate measures do not align with each other, and the exact role of fourfold patterns is ambiguous. All in all, this raises the question about the overall balance of evidence that the original CPT paper ultimately provided in favor of, or against its own theory. Similarly, moving beyond the 1992 paper, and considering CPT's entire "functional menagerie" (Stott, 2006) of potential utility and weighting functions, it is unclear to us whether these modifications make the theory's scope or parsimony any less ambiguous and in what way. We are not aware of any consensus in the field as to how to weigh the many versions of the theory against each other and against competing theories, or how to determine which version of the theory provides unambiguously the best tradeoff between scope and parsimony, across all possible stimuli and tasks.

The primary purpose of this note is not to question the validity of CPT as a theory, nor to provide further grounds to endorse it, but rather, to call attention to the ambiguity of the evidence provided by the authors in support of their own theory. Nor is our goal to single out Tversky and Kahneman for a practice that appears rather wide-spread. Our goal is conceptual: How should we really think of theoretical scope in Psychology? History has been repeating itself, in that much behavioral decision research turns CPT's limitations against it. Scientific cost-benefit analysis in decision science, all too often, appears to focus on highlighting the cost of others' theories and the benefits of one's own proposals. The resulting fault lines have left us with various 'camps:' Some endorse Expected Utility Theory as their preferred theoretical idealization. Others consider CPT as their sweet spot for a theory of risky choice. Meanwhile, countless papers expand or modify prior theories in order to accommodate behavioral (ir)regularities that have been reported as evidence against those theories. Over time, each new generation has tended to develop its new proposals by calling out limitations in previous ideas[12], with little attention to the limitations of the new theses (for examples of notable exceptions, see Brandstätter et al.,

[12] We refrain from citing what are hundreds of papers.

2006; Brandstätter, Gigerenzer, & Hertwig, 2008; Loomes, 2010, who acknowledged and specified limitations of their own models) and, perhaps more importantly, with little discussion about what limitations are acceptable or inacceptable. The literature in risky choice often appears to follow a three-pronged research strategy: 1) Scrutinize the old theory by showing certain weaknesses and promote the new theory by showing that it overcomes that particular set of weaknesses. 2) Leave it to others to explore the new proposed theory's weaknesses. 3) Either ignore the other camps or defend the new theory vigorously against their challenges. We see little effort towards reconciliation among schools of thought that highlight different aspects of and approaches to decision making. We also detect little effort to weigh strengths and weaknesses of competing theories in a comprehensive manner.

While our discussion centered around decision making, our conclusions apply to the discipline more widely. In our view, Psychology should move from using theoretical scope primarily as a bludgeon to attack others' theories, and proceed towards pursuing more constructive goals. Every scientific theory has some limitations, especially in psychology. Behavioral scientists should make every effort, when proposing a new idea, to spell out the intended scope of this new theory. Proposals for new theories are far more interesting when they also delineate what would constitute critical tests, what would qualify as refutation of a new proposal, what is considered beyond a theory's intended scope, who the theory applies to when, where, and why. Likewise, more adversarial collaborations between 'camps' would help bring sense to the balkanized landscape of entrenched schools of thought (see, e.g., Table 1 of Mellers, Hertwig, & Kahneman, 2001, for useful guidelines on how to run such projects).

So, what are we to make of the fact that half of Tversky and Kahneman's participants in one of their studies appear to have violated their own theory on one stimulus? What are we to make of the internal inconsistencies among reported findings within one paper? It is unclear how to weigh the evidence in favor of CPT on the one hand and the limitations of CPT on the other hand, against the corresponding strengths and weaknesses of competing theories. Statistical Science actively researches and studies the trade-off in complexity and parsimony in statistical models, by counting parameters and degrees of freedom, by computing heuristic model selection indices like AIC and BIC, as well as by applying quantitative model selection tools such as Bayes factors. As Yarkoni (2020) argues in somewhat different words, the 'sampling' of stimuli, participants, and design features of psychological research effectively hides uncounted degrees of freedom in the data. Psychology still needs to properly define theoretical scope and theoretical parsimony beyond post-hoc statistical models. The discipline should move beyond weighing, say, 'good' and 'bad' stimuli or study designs heuristically, and develop methods, concepts, and standards for weighing theoretical scope against scientific simplicity. A first step is for scholars of different schools of thought to cooperate more systematically in synergizing the strengths of different theories. A second and easy-to-implement step is for scholars to specify what they mean by "diagnostic stimuli" and/or "critical tests," not only for existing theory, but also for their own proposed new theory. A third step is for scholars to be more cognizant that every behavioral theory has limitations, and therefore spell out, as explicitly as they can, what scope they envision for their proposed theory.

References

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

Bogdan, P., Cervantes, V., & Regenwetter, M. (n.d.). *Bridging the model-theory gap in mediation analysis: What does mediation reveal about individual people?* (Manuscript under review)

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*, 756–766.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The Priority Heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409-432.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008). Risky choice with heuristics: Reply to birnbaum (2008), johnson, schulte-mecklenbeck, and willemsen (2008), and rieger and wang (2008). *Psychological Review*, *115*, 281-290.

Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision*, *1*, 252-274.

Busemeyer, J., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1561.

Chen, M., Regenwetter, M., & Davis-Stober, C. (2021). Collective choice may tell nothing about anyone's individual preferences. *Decision Analysis*, *18*, 1-24.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*, 997-1003.

Davis-Stober, C., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review*, *126*, 865-879.

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., Ert, E., Roth, A., Haruvy, E., Herzog, S., Hau, R., … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15-47.

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, *21*, 575-597.

Falmagne, J.-C., & Doble, C. (2016). *On meaningful scientific laws*. Springer-Verlag.

Falmagne, J.-C., & Narens, A. (1983). Scales and meaningfulness of quantitative laws. *Synthese*, *55*, 287 - 325.

Grahek, I., Schaller, M., & Tackett, J. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*, 803-815.

Guest, O., & Martin, A. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*, 789-802.

Heck, D. (2021). Assessing the 'paradox' of converging evidence by modeling the joint distribution of individual differences: Comment on Davis-Stober and Regenwetter. *Psychological Review*. (in press)

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, *16*, 844-853.

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

Kellen, D. (2020, June). *The limited value of replicating classic patterns of prospect theory.* Retrieved July, 2, 2020, from https://go.nature.com/2YqwWXR

Kellen, D., Davis-Stober, C., Dunn, J., & Kalish, M. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, *16*, 767-778.

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126 - 138.

Lakens, D., & DeBruine, L. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science*, *4*(2).

Lleras, A., Zang, Z., Ng, G., Ballew, K., Xu, J., & Buetti, S. (2020). A target-contrast signal theory of parallel processing in goal-directed search. *Attention, Perception, & Psychophysics*, *82*, 394-425.

Loomes, G. (2010). Modeling choice and valuation in decision experiments. *Psychological Review*, *117*, 902-924.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? an exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275.

Mistler, S. (2012). Planning your analyses: Advice for avoiding analysis problems in your research. *Psychological Science Agenda*, *26*(11).

Moore, D. (2016). Preregister if you want to. *American Psychologist*, *71*(3), 238.

Murphy, R., & ten Brincke, R. (2018). Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates. *Management Science*, *64*, 308-326.

Narens, L. (2002). *Theories of meaningfulness.* Lawrence Erlbaum Associates.

Narens, L. (2007). *Introduction to the theories of measurement and meaningfulness and the use of invariance in science.* Lawrence Erlbaum Associates.

Navarro, D. J. (2021). If mathematical psychology did not exist we would need to invent it: A case study in cumulative theoretical development. *Perspectives on Psychological science*, *16*, 707-716.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for Cumulative Prospect theory. *Journal of Mathematical Psychology*, *55*, 84 - 93.

Nosek, B., Spies, J., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618.

Pachur, T., Schulte-Mecklenberg, M., Murphy, R., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology: General*, *147*, 147-169.

Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Pitt, M., Kim, W., & Myung, I.-J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, *10*(1), 29–44.

Popov, V., & Reder, L. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1-46.

Regenwetter, M., Cavagnaro, D., Popova, A., Guo, Y., Zwilling, C., Lim, S., & Stevens, J. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision*, *5*, 63-94.

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review*, *119*(2), 408-416.

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., . . .

Messner, W. (2014). QTest: quantitative testing of theories of binary choice. *Decision*, *1*(1), 2-34.

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

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

Regenwetter, M., & Robinson, M. (2019b). Tutorial: Nuisance or substance? Leveraging heterogeneity of preference. *The Spanish Journal of Psychology*, *22*, e60.

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

Roberts, F. (1985). Applications of the theory of meaningfulness to psychology. *Journal of Mathematical Psychology*, *29*, 311 - 332.

Roberts, F. (1998). Meaningless statements. In R. L. Graham, J. Kratochvil, J. Nesetril, & F. S. Roberts (Eds.), *The future of discrete mathematics.* Providence, RI: American Mathematical Society.

Roberts, F., & Rosenbaum, Z. (1986). Scale type, meaningfulness and the possible psychophysical laws. *Mathematical Social Sciences*, *12*, 77 - 95.

Robinaugh, D., Haslbeck, D., Oisín, R., Fried, E., & Waldorp, L. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725-743.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*, 944-954.

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to access the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, *22*, 391-407.

Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100.

Schneegans, S., Taylor, R., & Bays, P. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(34), 20959–20968.

Sherpard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1217–1323.

Simmons, J., Nelson, J., & Simonsohn, U. (2021). Pre-registration: why and how. *Journal of Consumer Psychology*, *31*(1), 151–162.

Simons, D. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

Singmann, H., Cox, G., Kellen, D., Chandramouli, S., Davis-Stober, C., Dunn, J., … Shiffrin, R. (2021). *Statistics in the service of science: Don't let the tail wag the dog.* (PsyArXiv)

Stott, H. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and*

Uncertainty, *32*, 101-130.

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, *16*, 717–724.

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*, 94—95.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.

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

van Frassen, B. (2008). *Scientific representation: Paradoxes of perspective.* New York: Oxford University Press.

van Rooij, I., & Baggio, G. (2020). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*, 682–697.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zwilling, C., Cavagnaro, D., Regenwetter, M., Lim, S., Fields, B., & Zhang, Y. (2019). QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference. *Journal of Mathetical Psychology*, *91*, 176-194.

Appendix: Proofs and Caveats about Inference

Two claims regarding the fourfold pattern study and their proof

In their 17 Gain lotteries with $p \geq 0.5$ reported in their Table 3, Tversky and Kahneman exclusively used prospects of the form $(X, p; Y, 1-p)$, with $X \in \{50, 100, 150, 200, 400\}$, $p \in \{0.5, 0.75, 0.9, 0.95, 0.99\}$, and either $Y = 0$ or $Y = \frac{X}{2}$. According to Equations 1 and 2, because $w^{+}(1) = 1$, the subjective value of receiving the expected value E of such a lottery for sure is E^{α}. We show that, in each of the above cases, CPT predicts preference for the lottery (except in the special cases where CPT either reduces to Expected Value theory or uses a constant utility).

Claim I: *For* $0.5 \leq p < 1$*,* $X \in \{50, 100, 150, 200, 400\}$*,* $Y \in \{0, \frac{X}{2}\}$*, and* $0 < \alpha \leq 1$*,*

$$w^{+}(p) \times X^{\alpha} + (1 - w^{+}(p)) \times Y^{\alpha} < (pX + (1-p)Y)^{\alpha}\,, \qquad (6)$$

unless $\alpha = \gamma = 1$*.*

Claim I’: *For* $0.5 \leq p < 1$, $X \in \{-50, -100, -150, -200, -400\}$, $Y \in \{0, \frac{X}{2}\}$, *and* $0 < \beta \leq 1$,

$$w^{-}(p) \times X^{\beta} + (1 - w^{-}(p)) \times Y^{\beta} > (pX + (1-p)Y)^{\beta}, \tag{7}$$

unless $\beta = \delta = 1$.

We only need to prove Claim I: Since α and β take the same values and since γ and δ likewise do, setting $\lambda = 1$ without loss of generality, it is straightforward that Inequality 7 follows from Inequality 6 after multiplying both sides by (-1). Hence Claim I’ follows directly from Claim I.

Proof of Claim I. We first show that $w^{+}(p)$ underweights large probabilities. For $\gamma = 1$, we already know that $w^{+}(p) = p$. We show that, $w^{+}(p) < p$, when $0.5 \leq p \leq 1$ and $0 < \gamma < 1$. Since $0 < p^{(\gamma-1)^2} < 1$, it holds that

$$p < p\frac{1}{p^{(\gamma-1)^2}} = p\frac{1}{p^{\gamma(\gamma-1)-(\gamma-1)}} = p\frac{p^{\gamma-1}}{p^{\gamma(\gamma-1)}} = \frac{p^{\gamma}}{p^{\gamma(\gamma-1)}}.$$

Since $0 < \frac{1-p}{p^{\gamma}} < 1$ when $0.5 \leq p \leq 1$, and since $\gamma - 1 < 0$, it holds that

$$1 - p < (1-p)\left(\frac{1-p}{p^{\gamma}}\right)^{\gamma-1} = (1-p)\frac{(1-p)^{\gamma-1}}{p^{\gamma(\gamma-1)}} = \frac{(1-p)^{\gamma}}{p^{\gamma(\gamma-1)}}.$$

Adding these together gives,

$$1 < \frac{p^{\gamma} + (1-p)^{\gamma}}{p^{\gamma(\gamma-1)}}.$$

Exponentiating each side by $\frac{1}{\gamma}$, taking inverses, and multiplying each side by p yields

$$p > p\frac{p^{\gamma-1}}{(p^{\gamma}+(1-p)^{\gamma})^{\frac{1}{\gamma}}} = w^{+}(p).$$

Since $w^{+}(p) < p < 1$, it also holds that $w^{+}(p) < p^{\alpha}$ for $0 < \alpha \leq 1$, and therefore

$$w^{+}(p) \times X^{\alpha} < p^{\alpha}X^{\alpha} = (pX)^{\alpha}\,.$$

This proves Inequality 6 when $Y = 0$ and $\gamma < 1$. When $Y = 0$ and $\gamma = 1$, the left side of Inequality 6 becomes pX^{α} and the right side becomes $(pX)^{\alpha}$. The function $g(\alpha) = p^{\alpha}$, for some fixed p with $0 < p < 1$ is decreasing in α. It takes its minimum value at $\alpha = 1$, namely $g(1) = p$. Hence, $pX^{\alpha} < p^{\alpha}X^{\alpha}$ when $\alpha < 1$, proving Inequality 6 when $Y = 0$, $\gamma = 0$, and $\alpha < 1$.

When $Y = \frac{X}{2}$ and $\alpha = 1$, the left side of Inequality 6 becomes $\frac{w^{+}(p)}{2}X + \frac{X}{2}$ and the right side becomes $\frac{p}{2}X + \frac{X}{2}$. Then, Inequality 6 holds, as long as $\gamma < 1$, since we have shown that $w^{+}(p) < p$ in that case. To prove Claim I for $Y = \frac{X}{2}$ when $0 < \alpha < 1$, we need to prove that

$$w^{+}(p) \times \left(X^{\alpha} - \left(\frac{X}{2}\right)^{\alpha}\right) + \left(\frac{X}{2}\right)^{\alpha} < \left(\frac{X}{2} + \frac{p}{2}X\right)^{\alpha}.$$

Since $w^{+}(p) < p$, and since $X^{\alpha} > 0$, it suffices to show that

$$p \times \left(1 - \left(\frac{1}{2}\right)^{\alpha}\right) + \left(\frac{1}{2}\right)^{\alpha} \quad < \quad \left(\frac{1+p}{2}\right)^{\alpha}, \text{or, in other words}$$

$$2^{\alpha}p + 1 - p \quad < \quad (1+p)^{\alpha}\,.$$

For any fixed α we consider the function $f(p, \alpha) = 2^{\alpha} p + 1 - p - (1+p)^{\alpha}$ with $0.5 \leq p < 1$. The second derivative with respect to p of this function is

$$\frac{\partial^2}{\partial p^2} = -\alpha(\alpha - 1)(p+1)^{\alpha - 2} > 0, \text{ with } 0 < \alpha < 1.$$

Since $f(0.5, \alpha) \leq 0$ and $f(1, \alpha) = 0$, and since f is strictly convex in p for this domain, $f(p, \alpha) < 0$. This completes the proof of Claim I.

Two claims regarding the loss aversion study and their proofs

Before we state and prove the remaining Claims II and III, we provide additional notation and basic insights. For $p = \frac{1}{2}$, the weighting function for CPT, given in Eq. 2, becomes

$$w^{+}\left(\frac{1}{2}\right) = 2^{-(\gamma + \frac{1}{\gamma} - 1)} \qquad (0 \leq \gamma \leq 1).$$

Henceforth, we label this quantity as w_{γ}. The derivative of w_{γ}, as a function of γ, is

$$\frac{-ln(2)(\gamma^2 - 1)2^{-(\gamma + \frac{1}{\gamma} - 1)}}{\gamma^2}.$$

This derivative is positive and w_{γ} attains its maximum value of 0.5 at $\gamma = 1$. We later need to consider the quantity $\frac{1 - w_{\gamma}}{w_{\gamma}}$. Its derivative is $\frac{-1}{w_{\gamma}^2}$, which is negative. Thus, $\frac{1 - w_{\gamma}}{w_{\gamma}}$ is a strictly decreasing function of both w_{γ} and of γ. The minimum of $\frac{1 - w_{\gamma}}{w_{\gamma}}$ equals 1, namely when $\gamma = 1$.

Claim II: *Cumulative Prospect Theory does not permit $x \leq 149$ in Problem 7 of Tversky and Kahneman's (1992) test of loss aversion. If we take Tversky and Kahneman's median x of 149 at face value, then at least half the population violates CPT in this decision problem.*

Proof. For Problem 7 (see Table 6 of Tversky and Kahneman (1992) and our Table 3) a person who satisfies CPT, and who is indifferent between a 50/50 chance of winning \$50 or \$120 and a 50/50 chance of winning \$20 or $\$x_7$, must have values of γ, α at which the following constraint holds

$$\frac{1-w_\gamma}{w_\gamma} = \frac{x_7^\alpha - 120^\alpha}{50^\alpha - 20^\alpha}. \tag{8}$$

In particular, whenever $x_7 = 149$, then

$$\frac{1-w_\gamma}{w_\gamma} = \frac{149^\alpha - 120^\alpha}{50^\alpha - 20^\alpha}. \tag{9}$$

The derivative of the right hand side, as a function of α, is positive. Therefore, $\frac{149^\alpha - 120^\alpha}{50^\alpha - 20^\alpha}$ is increasing in α and it attains the maximum value of 0.966 at $\alpha = 1$. Since the left hand side of Equation 5 is bounded from below by 1 and the right hand side is bounded from above by 0.966, the equality cannot hold. Furthermore, replacing x_7 by values smaller than 149 will further reduce the right hand side and aggravate the discrepancy. Hence, Equation 4 cannot hold for values of 149 or smaller. This completes the proof of Claim II.

Claim III: *For an odd number of decision makers, it is not possible to satisfy*

Cumulative Prospect Theory with median $\gamma = 0.61$ and median $x_8 = 401$ in Problem 8 of Experiment 2 in Tversky and Kahneman (1992). Therefore, Tversky and Kahneman's median γ is incompatible with their median x_8 among their 25 decision makers.

Proof. For Problem 8, a person who is indifferent between a 50/50 chance of winning \$100 or \$300 and a 50/50 chance of winning \$25 or \$$x_8$, must have values of γ, α at which the following constraint holds

$$\frac{1 - w_\gamma}{w_\gamma} = \frac{x_8^\alpha - 300^\alpha}{100^\alpha - 25^\alpha}.$$

In particular, if $x_8 = 401$, then

$$\frac{1 - w_\gamma}{w_\gamma} = \frac{401^\alpha - 300^\alpha}{100^\alpha - 25^\alpha}. \tag{10}$$

The derivative of the right hand side, as a function of α, is positive. Therefore, $\frac{401^\alpha - 300^\alpha}{100^\alpha - 25^\alpha}$ is increasing in α and it attains the maximum value of 1.347 at $\alpha = 1$. This corresponds to $w_\gamma = 0.426$ and, in turn, $\gamma = 0.621$. Since the left hand side of Equation 6 increases as γ decreases and since the right hand side is at its maximum when $\gamma = 0.621$, therefore, regardless of the value of α, the equality in Equation 6 cannot hold when $\gamma < 0.621$. As a consequence, replacing x_8 by values smaller than 401 will further reduce the right hand side and require $\gamma > 0.621$.

For an odd number N of decision makers, a median x_8 value of 401 means that $\frac{N+1}{2}$

many must have a value of $x_8 \leq 401$, and hence a value of $\gamma \geq 0.621$. Yet, in contradiction to that requirement, a median γ value of 0.61 means that $\frac{N+1}{2}$ many must have a value of $\gamma \leq 0.61$. This completes the proof of Claim III.

Caveats about inference:

Intuitively, 0.966 does not look very different from 1 (in the Proof of Claim II), and, likewise, 0.621 does not look very different from 0.61 (in the Proof of Claim III). If we were to treat all medians and/or parameter estimates as outcomes of random variables due to random sampling of respondents and responses, we could ask whether fewer than half of the population may have violated CPT in Problem 7 and estimate how many people are exceptions to CPT according to that decision problem. We could also ask whether an estimated $\gamma = 0.621$ is statistically significantly larger than an estimated $\gamma = 0.61$, etc. However, short of having a theory of individual differences that would constrain or structure the distributions of observable values of x and estimated γ values (hence also of w_γ), such a test appears to be either ill-defined or unavailable.