Two steps to risk sensitivity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distributional reinforcement learning (RL) – in which agents learn about all the possible long-term consequences of their actions, and not just the expected value – is of great recent interest. One of the most important affordances of a distributional view is to facilitate a modern, measured, approach to risk when outcomes are not completely certain. By contrast, psychological and neuroscientific investigations into decision making under risk have utilized a variety of more venerable theoretical models such as prospect theory that lack axiomatically desirable properties such as coherence. Here, we consider a particularly relevant risk measure for modeling human and animal planning, called conditional value-at-risk (CVaR), which quantifies worst-case outcomes (e.g., vehicle accidents or predation). We first adopt a conventional distributional approach to CVaR in a sequential setting and suggest that many human decision-makers in the well-known two-step task are substantially risk-averse – leading to behavior that had previously been accounted for by appealing to stickiness or perseveration. We then consider a further critical property of risk sensitivity, namely time consistency, showing alternatives to this form of CVaR that have desirable characteristics of their own. We use simulations to examine settings in which the various forms differ in ways that have implications for human and animal planning and behavior.

1 Introduction

Risk is integral to decision making, arising whenever there are uncertain outcomes. It is especially critical when those outcomes are potentially calamitous, and plays an important role in psychiatric illness [1]. However, psychological investigations into choice under risk (i) has yet to embrace the strong formal foundations being developed in finance, AI and machine learning; and (ii) has mostly focused on one-shot decision tasks, despite the ubiquity of situations in the real-world that require planning, and growing interest in multi-step tasks for elucidating richer mechanisms of choice [2, 3].

To address these lacunæ, we consider a modern, theoretically-well founded coherent [4] risk measure called conditional value-at-risk (CVaR). This exactly quantifies the lower tail of possible outcomes – those which are important for survival, and also tend to drive our most persistent worries. CVaR has been applied to sequential decision problems, notably by means of distributional reinforcement learning [RL; 5–7]. In this paradigm, the agent (or decision maker) estimates whole distributions for potential outcomes that can arise from its actions. Risk measures, such as CVaR, can be applied to these distributions to adjust decision-making to any desired level of risk sensitivity.

In this paper, we combine CVaR with a distributional approach to examine risk sensitivity in the intensively-investigated two-step sequential decision task [2]. We find that the choices of many individuals reflect a large degree of risk aversion. Moreover, we show that more standard analyses ascribe this to enhanced choice perseveration and/or reduced estimates for learning rate, thus misspecifying the effects of relatively high levels of uncertainty. Incorporating risk sensitivity into choice, through CVaR or other risk measures, however, raises subtle issues regarding the consistency

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

of choices over time [8–11]. Such issues are well explored in psychology in the context of hyperbolic temporal discounting [12], but their thorough investigation in finance and operations research has yet to permeate judgement and decision-making. We discuss how a direct incorporation of CVaR into distributional RL can lead to time inconsistency and point to two additional time-consistent approaches to CVaR for sequential decision making. Finally, we show how these three approaches can give rise to distinct patterns of behavior in a problem designed to tease them apart and which can serve as the basis for future empirical investigation.

Conditional value-at-risk (CVaR)

A risk measure $\rho(Z)$ maps a probability distribution of outcomes associated with a random variable Z to a real number. In distributional RL, Z is typically a discounted sum of rewards minus costs (i.e. the return). $\rho(Z)$ represents the risk inherent in the uncertainty about Z, and is often interpreted as the amount one is willing to pay to avoid adopting this risk.

Well known risk measures include the variance and the α%-quantile of the distribution (called value-at-risk or VaR_α, defined as $\text{VaR}_\alpha(Z) = F^{-1}(\alpha) := \inf\{z : F(z) \geq \alpha\}$, for cumulative distribution function $F(z)$; see Figure 1a). Human reasoning is often modelled using cumulative prospect theory [13, 14], mean-variance [15, 16], or mean-variance-skew characterizations [17, 18]. However, while venerable, none satisfies the full set of axiomatically desirable properties associated with *coherent risk measures* [4]. For instance, the VaR is not sub-additive (so fails to reward diversification), and variance is neither sub-additive nor positively homogeneous (as it changes non-linearly with the units in which costs or rewards are measured). It is thus important to consider other possibilities.

Conditional value-at-risk (CVaR) was introduced as a coherent risk measure [4], and is particularly popular in finance, robotics, operations research, and recently machine learning and RL [6, 19–25]. For the lower tail of a continuous distribution, it is defined as the average of the values lower than the VaR (Figure 1a):[1]

$$\text{CVaR}_\alpha[Z] := E[Z|Z < \text{VaR}_\alpha(Z)] \tag{1}$$

α determines risk aversion by emphasizing the lower tail of the distribution more or less (Figure 1b).

That CVaR concentrates on the lower tail makes it particularly attractive for capturing aspects of normal and pathological reasoning in animals, whose lives often hang on rather thin threads, and humans, who can catastrophize about diverse, unlikely, possibilities. But it is then particularly important to consider CVaR in sequential decision-making, since this represents the ecological norm; and is etched into the neural structure of decision-making. We therefore start by examining CVaR in the simplest such problem that is widely studied in humans, the two-step task.

One critical issue for CVaR is time consistency – outcomes that constitute the tail of a distribution of possibilities at a state may no longer even be possible following a stochastic transition from that state. What, then, does it mean for the CVaR at the two states to be consistent? We consider two additional versions of CVaR that approach time consistency in radically different ways, and use simulations to illustrate the differences between all three.

2 Modeling risk sensitivity in human planning: CVaR in the two-step task

The two-step task (Figure 2) is a popular option for studying sequential choice in humans (and animals) [2]. It was originally designed to investigate model-based (MB) and model-free (MF) learning and planning, distinguishing the two by examining the consequences of progressive changes in the probabilities of rewards. However, these changes necessarily induce uncertainty – which could affect risk-sensitive subjects. We therefore used a CVaR-based form of MB and MF reasoning to fit a very substantial dataset of human behavior in this task [out of more than 2000 participants, the 792 who responded on every trial in ref. 3].

Task: On each of 200 trials, participants make decisions at two successive stages or steps. At the first stage, participants choose between two actions (the fractals in Figure 2a). As shown by the arrows in the figure, depending on their choice, they then transition to one of two possible second

[1] For discrete random variables, alternative definitions such as $\text{CVaR}_\alpha(Z) := \sup_\nu\{\nu - \frac{1}{\alpha}E[(\nu - Z)^+]\}$ [26] are used.

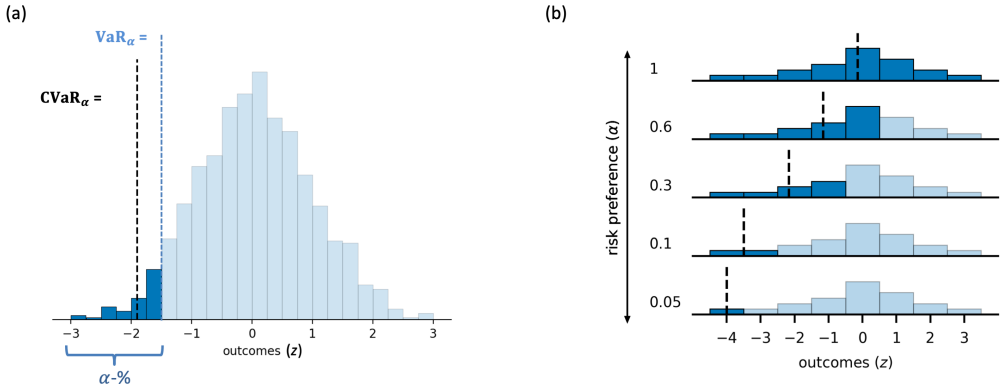

Figure 1: Conditional value-at-risk (CVaR). (a) CVaR$_\alpha$ is the average of the values in the lower α-% of a distribution, i.e. below the VaR$_\alpha$. (b) Adjusting α emphasizes the lower tail of the distribution more or less and sets the level of risk aversion. At $\alpha = 1.0$, CVaR is equal to the mean. As α decreases to 0, it approaches the minimum.

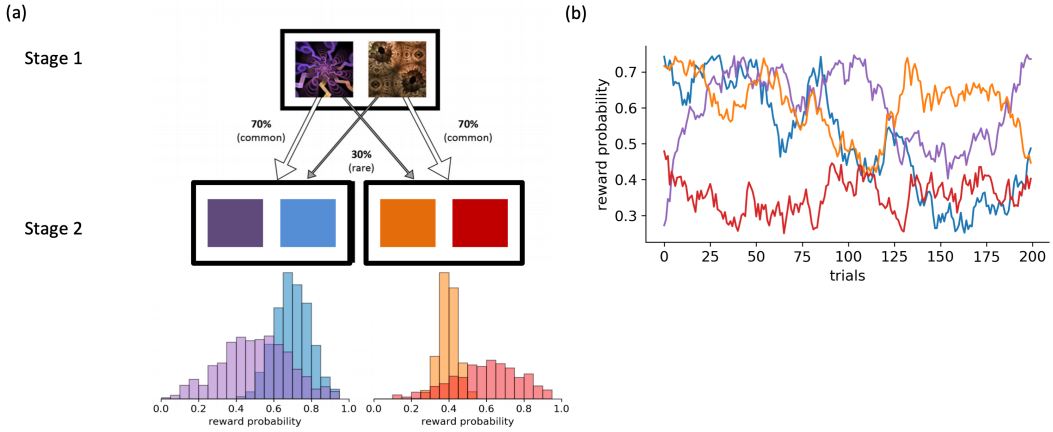

Figure 2: Two-step task. (a) On each trial, participants choose between two options (depicted as fractals) in stage 1 and then transition stochastically to one of two states (blue/purple or orange/red) in stage 2. They again choose between two options (the colored squares) and receive a binary reward. (b) This reward is sampled according to a probability that drifts randomly across trials (separately for each option). Distributions over reward probabilities are assumed by the model to be estimated for each option (depicted in panel a).

stage states either 70% or 30% of the time. Each second stage state has its own pair of options (the colored squares) between which the participants must then choose. According to this second choice, the participant receives a binary outcome, which is drawn according to a probability that drifts randomly across the trials (shown by color in Figure 2b).

Model: We included CVaR in a model that is conventionally applied to the two-step task. We assume that participants learn a distribution of values [5] associated with each of the four second-stage options using an approximate Kalman filter:[2]

$$\mu_{t+1} = \mu_t + \lambda[r_t - \mu_t] \qquad \Psi^2_{t+1} = (1-\phi^2)\Psi^2_t + \eta^2 - \lambda\Psi^2_t \tag{2}$$

updating the mean μ_t of each distribution on each trial using the observed outcome r_t via a delta-rule with a participant-specific learning rate λ, and also updating the variance Ψ_t. When the outcome is not observed (i.e. when the participant chose a different option), the mean is updated towards 0.5 and the variance is updated without the last term in equation (2). Thus, the dispersion parameter η^2

[2]We approximate the uncertainty using Gaussian statistics. Since rewards are binary, we also tried a forgetful beta-binomial model, but this fit participants' data less well.

controls the increase of the variance whether or not outcomes are observed, and the learning rate λ controls the amount it decreases when they are. The term $(1-\phi^2)$ controls the asymptotic variance for unobserved outcomes (in relation to η^2) and was set based on the other two parameters (see later).

The $\text{CVaR}_{\alpha,t}(a)$ at risk preference α (estimated per participant) was calculated for each option a using the mean $\mu_t(a)$ and variance $\Psi_t(a)$ on each trial t. Participants' second stage choices were modeled using a soft-max choice rule $P(2^{nd}\text{ stage choice}=a) \propto \exp(-\tau^{2nd}\text{CVaR}_{\alpha,t}(a))$, where the parameter τ^{2nd} controls the relative stochasticity/determinism.

Decisions between the two first-stage options were modeled as involving a combination of model-free (MF) and model-based (MB) approaches to value estimation – both of which were modified to include CVaR. MF estimates were calculated using the same formulæ, learning rate λ and dispersion parameter η^2 as at the second stage to learn an additional pair of means and variances based on the actual outcomes received in the second stage.[3] For model-based estimation, the 70% / 30% transition probabilities were used to calculate a mixture distribution for each of the top stage actions, whence the CVaR was calculated directly. $\text{CVaR}_{\alpha,t}(a)$ for the model-free and model-based first stage distributions again determined the choice probability through a soft-max choice rule:

$$P(1^{st}\text{stage choice}=a) \propto \exp(\tau^{MB}\text{CVaR}^{MB}_{\alpha,t}(a) + \tau^{MF}\text{CVaR}^{MF}_{\alpha,t}(a) + \tau^{sticky}\delta_{a,a_{t-1}}) \qquad (3)$$

Parameter τ^{sticky} was included to capture the tendency of participants to repeat (or to switch) the previously chosen action regardless of its value.

The 7 parameters of the final model were estimated for each participant: CVaR-based risk-sensitivity $\alpha \in [0.1, 1]$, learning rate $\lambda \in [0, 1]$, dispersion $\eta^2 \in [0.001, 0.1]$, perseveration $\tau^{sticky} \in [0, 50]$ and three inverse temperature parameters $\tau^{2nd}, \tau^{MB}, \tau^{MF} \in [0, 50]$.

Parameters were estimated in Python using L-BFGS-B. Parameter recovery simulations suggested that estimating both η^2 and $(1-\phi^2)$ was difficult, so we determined the value of $(1-\phi^2)$ that would pin the asymptotic variance to 0.1. This was done so that a never-chosen option with an estimated mean 0.5 would have a 10% CVaR of 0 (which is appropriate since the outcomes themselves were between 0 and 1 and we imposed a lower bound of 0.1 on α). The learning rate lay between 0 to 1 and was additionally constrained such that $\lambda < (1-\phi^2)$.

Importantly, setting $\alpha = 1$ arranges for CVaR choices that depend only on expected values (in which case the variance estimation can also be removed, because it does not influence choice probability). This mean-model, which we contrast with the $\text{CVaR}_{\alpha<1}$, is very similar to typical models used in the two-step task [2, 3], but is more directly comparable to its risk-sensitive counterpart.[4]

Results

We first compared the the mean-model $\text{CVaR}_{\alpha=1}$ to the risk-sensitive $\text{CVaR}_{\alpha<1}$. Using BIC to account for the two extra parameters, we found that the mean- and risk-sensitive models were roughly equally preferred overall. However, the red points in Figure 2a shows that many subjects fit substantially better when allowed $\alpha < 1$; Figure 2b duly shows that many of those had $\alpha < 0.2$, indicating substantial aversion to risk.[5]

To investigate the improved fit and for insight into the choice characteristics associated with risk aversion, we analyzed how other parameters altered from the $\alpha=1$ case for significantly risk-averse subjects. There were two, related, changes: participants' estimated learning rates were higher in the CVaR-model (Figure 3c), and their estimated perseveration parameters were lower (Figure 3d). One potential reason for these is that participants tended *not* to prefer less frequently chosen, and thus more uncertain options, even when the more certain option was apparently worse.

To demonstrate this directly, we simulated choices from the CVaR-model at increasing levels of risk sensitivity ($\alpha \in [1.0, 0.6, 0.3, 0.1]$) in response to a predetermined set of outcomes, with the

[3] In some models of the two-step task, model-free values are also updated using the values of the second stage, however we found that this did not improve the fit of the model, so we omit it for the sake of parsimony.

[4] We found that the $\alpha = 1$ fits just as well as the original model employed by [3].

[5] To assess the estimability/identifiability of these parameters, we generated new data based on participant's parameters and re-estimated the model; a rank-correlation of 0.71 between the generative and recovered CVaR α parameter indicated moderately good estimability/identifiability; Supplemental Figure 1.

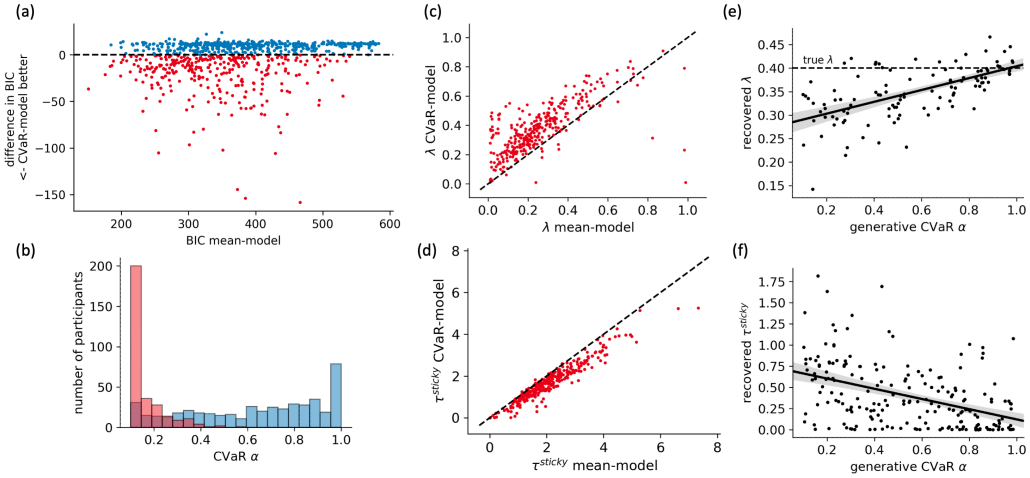

Figure 3: Two-step task modeling results. (a) Many participants are better fit by the CVaR-model than the mean-model, and (b) are substantially risk averse, as estimated by the CVaR α parameter. (c-d) Including risk sensitivity increases/decreases estimates of learning rate and perseveration, respectively, in these participants. (e-f) Simulations show that fitting a risk-agnostic model to risk-sensitive behavior can lead to underestimated learning rates (e) and/or apparent perseveration (f). Shading around the regression line in panels e-f indicate 95%-confidence intervals. Note that 3 outliers located near (x=12,y=10) were removed from panel d.

other parameters held constant (Figure 4). Option A is chosen by design for the first 6 trials; but the model is then allowed to switch to an option B after observing what it thinks is a sufficient number of negative outcomes. For $\alpha = 1$, this switch occurs when the mean estimate for option A crosses that estimate for option B at 0.5 (trial 10 in panel b). For $\alpha < 1$ (panels c-e), this switch occurs when the $\text{CVaR}_{\alpha<1.0}$ estimate for the options cross. Due to the uncertainty around the mean and the fact that the model is more uncertain about option B than A, the crossing point occurs later for lower α (i.e. on trials 11, 12, and 13). Thus, what looks like a tendency to stick with an apparently worse option, here arises from an aversion to uncertainty. For $\alpha = 1$, it can only occur if the learning rate is low (so the value of the chosen option does not decrease too much upon experiencing an unfortunate outcome), and/or with a high perseveration parameter. But, as a consequence, genuine risk aversion (or uncertainty aversion) might be misattributed to one of these parameters, something we show in Figure 3e (where inferred learning rates are too low for low α) and Figure 3f (where a non-perseverative CVaR model is fitted with substantial perseveration for low α). Note, though, that in participants' data, some perseveration remains even with $\text{CVaR}_{\alpha<1}$, as setting $\tau^{sticky} = 0$ led to worse BIC values.

3 Three approaches to sequential risk and time (in)consistency

One of the most important issues for sequential decision-making is time-consistency – that the choices the decision maker at time t assumes will be executed at time $t + 1$ are indeed carried out.[6] Inconsistency, famously caused by hyperbolic discounting [12], can lead to potentially sub-optimal, expensive, commitment behavior. Another route to time inconsistency comes from mis-handling risk sensitivity [8–11, 28–36] – and, indeed, different variants of CVaR approach this issue in different ways. The two-step task lacks sufficient temporal complexity to make this issue clear. Therefore in this section we examine it in greater depth, and, using simulations, suggest a task that highlights the possibilities.

[6] An intimately related notion of time consistency involves the consistency of successive risk evaluations of a stochastic process rather than choices per se. Of course, this then can lead to inconsistent choices. See [27–29] for a more in-depth discussion.

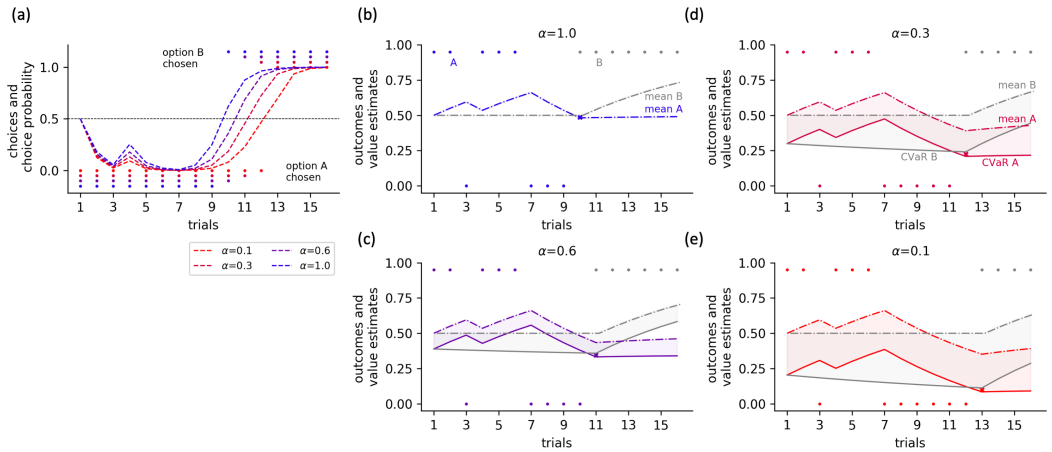

Figure 4: CVaR-based perseveration. (a) The CVaR-model is simulated at four values of α on a series of choices between options A and B. For low α (high risk sensitivity), the model chooses option A for more trials, despite receiving 0 outcomes after trial 7; binary outcomes are shown as dots in panels b-e. (b-e) The switch from A to B occurs when the model's estimate for the CVaR (or mean for $\alpha = 1.0$) of A (in color) goes below that of B (in gray). The crossing point occurs later for lower levels of α due to uncertainty around the mean which is greater for B.

Fixed, precommitted, and nested CVaR

In the two-step task, we followed recent work on distributional RL [6] in assuming that participants applied CVaR at the same risk preference (α) to the estimated value distributions at both the first and second stage. We call this the fixed approach or fCVaR. To see the issue with time consistency, consider calculating the optimal fCVaR$_{\alpha=0.1}$ policy for the simplified two-stage choice scenario depicted in Figure (5)a.

From the perspective of the top state A, consider the two policies $\pi = \{a_1 = \textit{right}, a_2 = \textit{right}\}$, which leads to a distribution of outcomes $p = \{0.1 : 0, 0.9 : 2\}$ and $\pi' = \{a_1 = \textit{right}, a_2 = \textit{left}\}$, which leads to a distribution of outcomes $p' = \{0.01 : -2, 0.09 : 1, 0.9 : 2\}$. Just considering these overall distributions, π' has an CVaR$_{0.1}$ of 0.7; whereas π has a CVaR$_{0.1}$ of 0, so π' would be preferred.

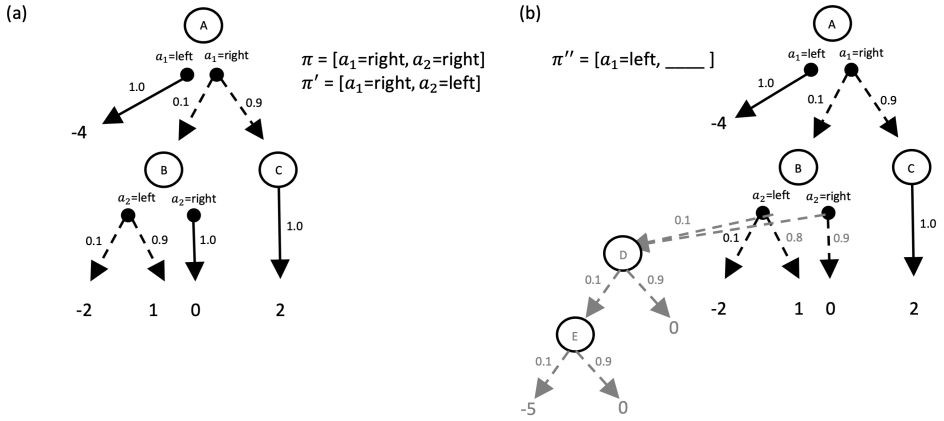

Figure 5: Three approaches to CVaR in a two-stage problem. (a) A choice of left or right can be made in both state A and B (if state B is visited). Stochastic transitions are denoted by dashed lines (with the probabilities shown), and the possible terminal outcomes are [-4, -2, 1, 0, 2]. Fixed and precommitted approaches to CVaR (discussed in the text) choose policy π and π', respectively. (b) In an adjusted problem, additional state transitions (colored in gray) are appended to state B. The nested approach to CVaR chooses policy π in the left problem, but π'' in the right problem.

However, if, having taken action $a_1 = \textit{right}$, the agent arrives at state B, now the $\text{CVaR}_{0.1}$ of $a_2 = \textit{left}$ is -2 and the $\text{CVaR}_{0.1}$ of $a_2 = \textit{right}$ is 0, so the agent will defect on its original plan π'.

One way to circumvent this issue with time-consistency is to skirt dynamic evaluation altogether and precommit to evaluating risk with respect to only one stage or time-point (the start being most natural) – thus enforcing π' at state A. One way to make this contract is to change the value of α after each transition in the light of the probability of its happening. For instance, the decision in π' to go *left* rather than *right* at state B was based on considering all of the outcomes in that state (i.e. the decision was based considering CVaR at $\alpha = 1.0$ rather than $= 0.1$). This dynamic adjustment of α prevents time inconsistency by allowing later risk evaluations and decisions all to be coordinated with respect to the single risk preference at the start stage [33].

A second way to deal with time-consistency is to evaluate risk dynamically using a series of nested one-step (conditional) risk measures [11, 37]:

$$\rho_{k,N}(R_k, \ldots, R_N) = R_k + \rho_k(R_{k+1} + \rho_{k+1}(R_{k+2} + \cdots + \rho_{N-2}(R_{N-1} + \rho_{N-1}(R_N)) \cdots)), \quad (4)$$

where R_k is the reward (or cost) for time step k.

If CVaR_α with the same value of α is used for each conditional risk measure ρ_k, this is known as nested or nCVaR_α [38]. Although like fCVaR, it keeps risk preferences fixed across time, it applies the CVaR at each stage to subsequent CVaR evaluations (which themselves are random *scalars*) rather than to the *full distribution* of future random costs or rewards under the policy lower in the tree.

One consequence of this nesting or compounding of risk evaluations, is that nCVaR can sometimes become much more conservative than the other two approaches. To see this, consider panel b in Figure 5. Here, the same decision tree is appended with a set of alternative transitions (depicted in gray) from state B. Now, each possible action is additionally associated with a 10% chance of ending up in state D, which either returns 0 with probability 0.9 or makes a transition to yet another state E with probability 0.1. The outcomes in this state are either -5 (the worst possible outcome) or again 0. Importantly, from the perspective of state B, the -5 is so remote (having a probability of merely 0.1^3) that it barely impacts the distribution there (and therefore pCVaR and fCVaR). However, due to the nested structure of nCVaR (at $\alpha = 0.1$, the -5 is propagated backwards first to E then to D, and then in fact, all the way back to the top state A. As a consequence, the nCVaR agent would then rather choose $\pi'' = \{a_1 = \textit{left}, \text{N/A}\}$, for a certain loss of -4.

Of course at $\alpha = 1.0$, all three approaches are equivalent to using the expected value, which is time-consistent. As α approaches 0, all three approaches again become equivalent, but to the worst-case risk measure [39], which is also time-consistent.

The three approaches in a gridworld

As noted, the two-step task is too simple to surface these issues (evaluating the other forms of CVaR leads to equivalent results on the 792 subjects). Thus, we used the simulations in Figure 6 to highlight their differences as the basis of possible future tests.

Consider an agent (or decision maker) who starts in a gridworld in the top left corner and can move either right or left. Exiting the gridworld on the right hand side informally represents a goal and is associated with a +3 reward, while exiting the map on the left hand side represents quitting and a loss of -2. The agents actions are stochastic with the possibility of moving downwards with some error probability. If the agent falls off the bottom of the gridworld, a substantial loss of -15 is incurred (schematized by the lavapit). Note the right action has twice the error probability as the left action, meaning that the heading towards the goal is riskier than attempting to quit; this leads to interesting differences for the three approaches.

Optimal polices for the pCVaR, fCVaR and nCVaR were calculated using dynamic programming equations (see Appendix B; based on [11, 22]) for various levels of α (for pCVaR, which requires interpolating between different values of alpha, we used 21 log spaced points, following [22]). The optimal policies for each approach to CVaR at a moderate-to-high level of risk (i.e. $\alpha = 0.18$) are depicted in Figure 6b.

For pCVaR (with a start state risk of $\alpha = 0.18$), the optimal policy in every state is to head towards the goal. This is the case even in the bottom row, where rightward actions are much more risky due to

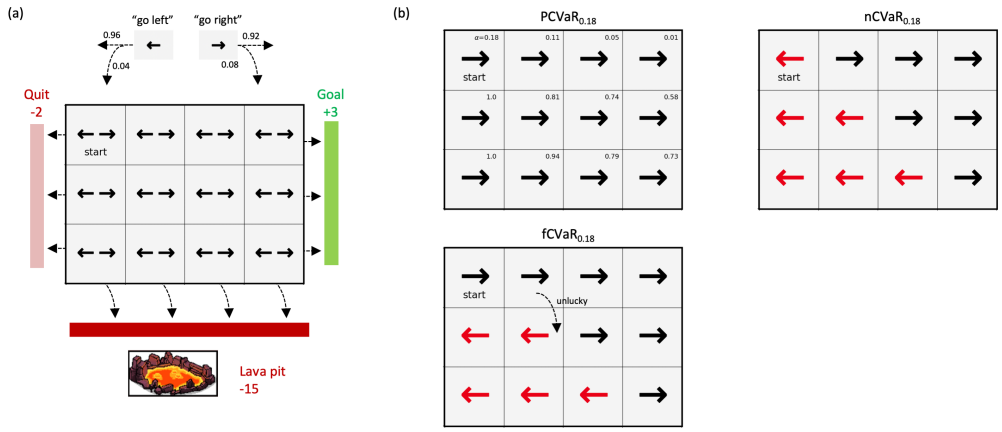

Figure 6: Three approaches to CVaR in a gridworld. (a) The agent starts in the upper left corner and can choose to go left or right in each state. Actions result in stochastic transitions (as depicted at the top). Exiting the map on the left, right, or bottom results in an outcome of -2, +3, -15. (b) The optimal policy is shown for each of three approaches, highlighting their distinct behaviors.

the proximity of the lava pit. In fact, if the agent started in the bottom row with the same $\alpha = 0.18$, it would head left in an attempt to quit. However, since the pCVaR approach coordinates risk with respect to the start state and these bottom states are part of the start state's lower tail, the pCVaR agent knows that it is better to make 'riskier' decisions here for the sake of its former self.[7] Indeed, doing so yields a higher start state CVaR than abandoning its plans. In contrast, abandoning the pursuit of the goal is exactly what the fCVaR agent does, as it re-evaluates risk using $\alpha = 0.18$ at every state; i.e., it heads towards the left in rows 2 and 3 after getting knocked off course. Similarly to fCVaR, the nCVaR agent also re-evaluates risk using $\alpha = 0.18$ at every state. However, it chooses to quit from the start due to the nesting of risk (risk evaluated on top of future risk evaluations) – even from the start, the chances of the distal lava pit looms larger than it does for the other two approaches (at the same nominal level of alpha). At lower and higher values of alpha, the behavior of the three approaches more closely aligns, either quitting from the start or pursuing the goal from every state.

4 Discussion and related work

We first showed that adopting the modern risk measure CVaR [4] in a form based on ideas from distributional RL [6] suggests that a large minority of healthy volunteer subjects exhibit quite significantly risk-averse behavior. This was true even in a simple two-step task for extremely low stakes and with only rather limited amounts of uncertainty. That this risk aversion masqueraded as enhanced perseveration and slowed learning is a reminder of the complexities of building models of behavior, and an invitation to consider risk in models of other common tasks. These effects of risk in behavior can be complemented by considerations of its effects in the sort of off-line planning that has recently been suggested to model rumination in anxiety disorders, as subjects struggle to find ways of mitigating potential future threats [40]. That distributional RL has recently been suggested as a way of understanding facets of the activity of major neural systems involved in processing affective outcomes [7] offers a highly attractive link to understanding risk aversion in animals [and also tempts us to consider the role of other neuromodulators that have been implicated in representing and processing uncertainty; 41].

Since pathological effects of uncertainty play a malign role in the effects of many psychiatric conditions [1], it would be of great interest to design tasks, for instance based on the gridworld of section 3, that decompose various of its different facets to help determine which might be responsible. Indeed, this task can be seen as a structurally rich form of a popular method for assessing impulsivity – the balloon analogue risk task (BART; [42]). Many potential variants would be of interest – for instance studying forms of precommitment by allowing subjects to buy a form of 'insurance' at

[7] This increase in risk sensitivity can be seen in the average values of the adjusted alphas in each state; shown in the upper right hand corner of each grid cell.

the outset against one or more unlucky downwards transitions. We could modify the task further with intermediate rewards and punishments to examine the observation that individuals only take on greater risk after a loss, if it is not yet realized (in our terms, before reaching one of the two sides of the grid) [43]. Then, we might expect precommitments to be abandoned and risk to be re-evaluated to the degree to which previous outcomes seem irreversible or the current situation sufficiently distinct from the one in which the committment was made. One could examine consistency between subjects' choices in the environment and their willingness to pay to protect themselves ahead of time. Furthermore, richer distributions of possible outcomes could be used to reverse engineer uncertainty calculations in the brain given stable risk preferences.

One particular facet of uncertainty that we did not explore is that, in some circumstances, it is possible to collect more information that reduces it to acceptable levels. Indeed, modern risk measures have recently been applied applied in settings with model uncertainty (i.e. POMDPs and Bayes adaptive MDPs, [44–47]). Information gathering, as explored in the sequential information sampling task [48] has been of particular interest in obsessive compulsive disorder, where subjects have been algorithmically modelled as exhibiting less urgency to make up their minds [49]; it would be interesting to model them computationally as being more risk averse. Similarly, if trust is seen as willingness to risk vulnerability to others [50, 51], then risk sensitivity could be an important factor in social pathologies such as borderline personality disorder.

We showed that fCVaR, although intuitive, is not time consistent. The other forms of precommitted pCVaR and nested nCVaR are, which provides them with more attractive formal properties. In fact, one can see fCVaR as living inbetween pCVaR and nCVaR (rather literally, in the problem shown in Figure 6b). There are other potential interpolants – for instance, cases in which the updating of α following lucky or unlucky transitions [which justifies a form of gambler's fallacy; 40] is incomplete (or perhaps asymmetric). Such incompetent calculations have been suggested as underlying psychiatric conditions themselves [52].

Finally, although we focused on CVaR, there are many other risk measures that satisfy the coherence axioms, and indeed other more stringent conditions. Adding the requirements of comontonicity and law invariance lead to the class of *distortion risk measures* [53–55] (or equivalently, *spectral risk measures* [56]), which apply a distortion function to cumulative probabilities. This allows them to be linked directly to the *dual theory* of choice [57], therefore inheriting an additional set of rational choice axioms (i.e. the axioms of expected utility theory, with an alternative independence axiom)[58]. CVaR itself is part of this class, and can be used as a basis to construct all other members [36, 55].

5 Broader Impact

Coherent risk measures, despite their widespread use in formal applications and favorable theory, have yet to permeate fully psychological and neuroscientific models of decision making, especially for sequential problems. We take early steps in this direction using a widely-recognized and commonly used experimental paradigm, and a psychologically relevant coherent risk measure, CVaR. In doing so, we highlight a key issue that can arise, namely time inconsistency, with a naive application of risk measures to sequential choice, and discuss two alternative (time-consistent) approaches. We expect that bringing awareness to this issue will lead to interesting future discussions and empirical tests in psychological and neuroscientific communities. We do not anticipate our research to selectively impact some groups at the expense of others. Our study had the limitation of not examining how subjects might compute these various forms of risk sensitivity in a neurobiologically credible manner, or showing that the risk aversion that we inferred from behaviour in the two-step task would generalize to other choices that the subjects might make.

Acknowledgements

The authors have no competing interests to disclose.

References

[1] Matthias Brand, Kirsten Labudda, and Hans J Markowitsch. Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks*, 19(8):1266–1276, 2006.

[2] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[3] Claire M Gillan, Michal Kosinski, Robert Whelan, Elizabeth A Phelps, and Nathaniel D Daw. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, 5:e11305, 2016.

[4] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[6] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[7] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

[8] Kang Boda and Jerzy A Filar. Time consistent dynamic risk measures. *Mathematical Methods of Operations Research*, 63(1):169–186, 2006.

[9] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, David Heath, and Hyejin Ku. Coherent multiperiod risk adjusted values and bellman's principle. *Annals of Operations Research*, 152(1):5–22, 2007.

[10] Alexander Shapiro. On a time consistency concept in risk averse multistage stochastic programming. *Operations Research Letters*, 37(3):143–147, 2009.

[11] Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

[12] George Ainslie and Nick Haslam. Hyperbolic discounting. 1992.

[13] Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, March 1979.

[14] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[15] Harry Markowitz. Portfolio selection*. *The Journal of Finance*, 7(1):77–91, 1952.

[16] Rakesh K Sarin and Martin Weber. Risk-value models. *European Journal of Operational Research*, 70(2):135–149, 1993.

[17] Mkael Symmonds, Nicholas D Wright, Dominik R Bach, and Raymond J Dolan. Deconstructing risk: separable encoding of variance and skewness in the brain. *Neuroimage*, 58(4):1139–1149, 2011.

[18] Charlene C Wu, Peter Bossaerts, and Brian Knutson. The affective impact of financial skewness on neural activity and choice. *PloS one*, 6(2):e16838, 2011.

[19] Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

[20] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[21] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[22] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[23] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[24] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[25] Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

[26] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[27] Michel De Lara and Vincent Leclère. Building up time-consistency for risk measures and dynamic optimization. *European Journal of Operational Research*, 249(1):177–187, 2016.

[28] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[29] Birgit Rudloff, Alexandre Street, and Davi M Valladão. Time consistency and risk averse dynamic decision models: Definition, interpretation and practical consequences. *European Journal of Operational Research*, 234(3):743–750, 2014.

[30] David Kreps and Evan Porteus. *Temporal von Neumann–Morgenstern and Induced Preferences*. 1979.

[31] Larry G Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003.

[32] Berend Roorda and Johannes M Schumacher. Time consistency conditions for acceptability measures, with an application to tail value at risk. *Insurance: Mathematics and Economics*, 40(2):209–230, 2007.

[33] Georg Ch Pflug and Alois Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2):682–699, 2016.

[34] Alois Pichler and Alexander Shapiro. Risk averse stochastic programming: time consistency and optimal stopping. *arXiv preprint arXiv:1808.10807*, 2018.

[35] Rouven Schur, Jochen Gönsch, and Michael Hassler. Time-consistent, risk-averse dynamic pricing. *European Journal of Operational Research*, 277(2):587–603, 2019.

[36] Anirudha Majumdar and Marco Pavone. How should a robot assess risk? towards an axiomatic theory of risk in robotics. In *Robotics Research*, pages 75–84. Springer, 2020.

[37] Andrzej Ruszczyński and Alexander Shapiro. Conditional risk mappings. *Mathematics of operations research*, 31(3):544–561, 2006.

[38] Georg Ch Pflug and Werner Romisch. *Modeling, measuring and managing risk*. World Scientific, 2007.

[39] Stefano P Coraluppi and Steven I Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999.

[40] Christopher Gagne and Peter Dayan. Peril, prudence and planning as risk, avoidance and worry. *PsyArXiv*, 2021.

[41] Angela J Yu and Peter Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, 2005.

[42] Carl W Lejuez, Jennifer P Read, Christopher W Kahler, Jerry B Richards, Susan E Ramsey, Gregory L Stuart, David R Strong, and Richard A Brown. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2):75, 2002.

[43] Alex Imas. The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, 106(8):2086–2109, 2016.

[44] Jingnan Fan and Andrzej Ruszczyński. Dynamic risk measures for finite-state partially observable markov decision problems. In *2015 Proceedings of the Conference on Control and its Applications*, pages 153–158. SIAM, 2015.

[45] Apoorva Sharma, James Harrison, Matthew Tsao, and Marco Pavone. Robust and adaptive planning under model uncertainty. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 410–418, 2019.

[46] Mohamadreza Ahmadi, Masahiro Ono, Michel D Ingham, Richard M Murray, and Aaron D Ames. Risk-averse planning under uncertainty. In *2020 American Control Conference (ACC)*, pages 3305–3312. IEEE, 2020.

[47] Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. *arXiv preprint arXiv:2102.05762*, 2021.

[48] Luke Clark, Trevor W Robbins, Karen D Ersche, and Barbara J Sahakian. Reflection impulsivity in current and former substance users. *Biological psychiatry*, 60(5):515–522, 2006.

[49] Tobias U Hauser, Michael Moutoussis, Peter Dayan, and Raymond J Dolan. Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. *Translational psychiatry*, 7(12):1–10, 2017.

[50] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[51] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.

[52] Quentin JM Huys, Marc Guitart-Masip, Raymond J Dolan, and Peter Dayan. Decision-theoretic psychiatry. *Clinical Psychological Science*, 3(3):400–421, 2015.

[53] Shaun Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1):71–92, 1996.

[54] Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, pages 15–36, 2000.

[55] Hans Föllmer and Alexander Schied. *Stochastic finance*. de Gruyter, 2016.

[56] Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

[57] Menahem E Yaari. The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, pages 95–115, 1987.

[58] Hans Peter Wächter and Thomas Mazzoni. Consistent modeling of risk averse behavior with spectral risk measures. *European Journal of Operational Research*, 229(2):487–495, 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]

(c) Did you discuss any potential negative societal impacts of your work? [N/A] As our work investigates the psychology of decision making under risk, we do not anticipate it to have negative societal impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Data from the two-step task was already made publicly available at (https://osf.io/usdgt/) by the authors of the original study [3]. The code used for the analyses/simulations will be made available upon request.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Critical details for the analyses and simulations were provided in the text. Additional information will be provided in the supplemental materials.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All analyses and simulations were run on a single, personal machine with limited compute.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [N/A] The data was made publicly available without license at https://osf.io/usdgt/

   (c) Did you include any new assets either in the supplemental material or as a URL? [No]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Since we are reanalyzing data from [3] we did not discuss this. Details about the consent process can be found in the original work.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Details about the data collection process can be found in the original work [3].

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] See previous answer.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] See previous answer.