Time to explore: Adaptation of exploration under time pressure

Charley M. Wu[1,2*], Eric Schulz[3], Timothy J. Pleskac[4], and Maarten Speekenbrink[5]

[1]Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany
[2]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[3]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics
[4]Department of Psychology, University of Kansas, Lawrence, KS
[5]Department of Experimental Psychology, University College London, London, UK
[*]charley.wu@uni-tuebingen.de

ABSTRACT

How does time pressure influence exploration and decision-making? We investigate this question using a within-subject design to manipulate decision time (limited vs. unlimited) and use a range of four-armed bandit tasks, designed to independently manipulate uncertainty and expected reward. With limited time, people have less opportunity to perform costly computations, thus shifting the cost-benefit balance of different exploration strategies. Through behavioral, reinforcement learning (RL), reaction time (RT), and evidence accumulation analyses, we show that time pressure changes how people explore and respond to uncertainty. Specifically, participants reduced their uncertainty-directed exploration under time pressure, were less value-directed, and repeated choices more often. Since our analyses relate uncertainty to slower responses and dampened evidence accumulation (i.e., drift rates), this demonstrates a resource-rational shift towards simpler, lower-cost strategies under time pressure. These results shed light on how people adapt their exploration and decision-making strategies to externally imposed cognitive constraints.

Introduction

We have all experienced the pressure of making decisions under limited time. For instance, choosing what to order at a restaurant while the waiter waits impatiently behind your shoulder. Or deciding which analyses to run as a paper submission deadline looms near. With less time to think, we have less opportunity to perform costly computations. But does time pressure merely make us more noisy as we deal with the speed-accuracy trade-off[1,2]? Or are we able to adapt our decision-making processes, to make the best use of our cognitive resources given external constraints on our computational capacity[3–6]?

Here, we are interested in the cognitive processes involved in navigating the exploration-exploitation dilemma[7–9], which plays a key role when learning through interactions with the environment, such as in reinforcement learning[10] (RL) problems. Should you exploit your usual menu option or should you explore something new? The usual option may yield a predictably rewarding outcome, but forgoes the opportunity of learning about other menu items. A new option could lead to either a pleasant or unpleasant surprise, but will likely be informative for future decisions and could improve future outcomes.

Since optimal solutions to the exploration-exploitation dilemma are generally unobtainable[11,12] except in limiting cases[13–15](e.g., infinite time horizons), there is great interest in understanding the strategies that humans use[16,17]. Empirical evidence from a variety of experiments[8,18–21] and real-world consumer data[22] suggest people use a mix of two strategies: random and directed exploration. *Random exploration* increases the diversity of choices by adding stochasticity to the agent's behavioral policy, instead of only maximizing expected value. If you have only ever tried a handful of items on the menu, then you might

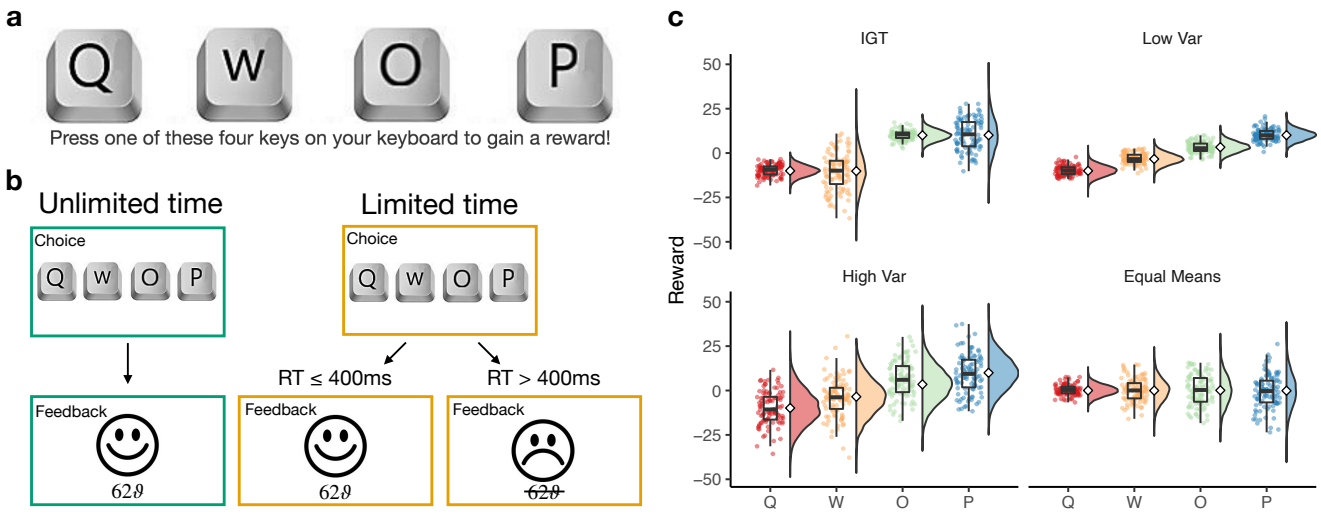

Figure 1. Experimental design. a) Time bandit task, where each option was randomly mapped to the $[Q, W, O, P]$ keys on the keyboard, with a different mapping each round. Participants completed 40 rounds (each containing 20 trials), where we manipulated time pressure (panel b) and payoff conditions (panel c) in a crossed, within-subject design. **b**) In *unlimited time* rounds, participants could take as long as they wanted to make each selection and received positive feedback (happy face) and were shown the value of the acquired payoff for 400ms. In *limited time* rounds, participants were only given 400 ms to make each selection. If they exceeded the time limit, they earned no rewards and received negative feedback (sad face) with the value of the payoff they could have earned crossed out. **c**) Each payoff condition specifies a normal payoff distribution for each option, with the means and variances described numerically in Table 1. The reward distributions are designed to compare how differences in reward expectations and differences in uncertainty influence choices (see Methods). Dots and the Tukey boxplots describe 100 randomly drawn payoffs, while the half violin plots show the generative distribution, with the diamond indicating the mean.

have an imperfect picture of what options are good. Thus, adding more variability to your choices may give you a better perspective about which options you should value. In contrast, *directed exploration* adds an exploration bonus to each option, proportional to the agent's level of uncertainty[23]. Rather than simply behaving more randomly, directed exploration is more strategic, prioritizing choices with the highest uncertainty to gain more information[24,25]. Perhaps there is an item on the menu you have never tried before. Directing your exploration to that novel item would be more effective at achieving an information maximization goal than choosing randomly, but may also incur more computational costs, since representations of uncertainty need to be factored into the decision-making process.

Limiting Decision Time

We manipulate decision time as a method for imposing external limitations on cognitive resources, to better understand the differential cognitive costs associated with random and directed exploration. With less time "budgeted" for costly computations, resource-rational decision makers[4,5] can be expected to choose cheaper strategies in order to achieve a better trade-off between the costs of computation and the benefits in terms of reward. One line of research on human decision-making commonly assumes that time pressure causes participants to rely more on "intuitive decision making"[26], making immediate outcomes more salient[27], and making people more reliant on fast, recognition-based processes than slower, more analytical processes[28]. Research using formal computational models has also related time pressure to changes in the speed-accuracy trade-off[29], yielding faster, less accurate decisions, but nevertheless still achieving an efficient rate of rewards[30,31]. However, there is disagreement in the literature about how time

pressure changes exploration patterns.

On the one hand, taxing cognitive capacities has been shown to *increase* exploration, producing less consistent and fewer expected value-maximizing decisions[32,33]. Similarly, people and monkeys placed under time pressure become more eager to select uncertain options, independent of outcome value[34,35]. Time pressure has also been linked to making people become more risk-seeking[36–38], although recent modeling work has challenged the reliability of this shift in risk preferences[33]. Nevertheless, a common thread is that limiting cognitive capacity reduces the scope or detail with which people evaluate different options[39–41], producing more impulsive decisions or a switch to simpler, heuristic decision-making strategies[42], both with similar patterns of increased exploration.

On the other hand, time pressure has also been shown to *decrease* exploration, leading to more repeat choice behavior and a reduced preference for uncertain options. Participants under time pressure are more likely to repeat previous actions[43], even to the detriment of producing more costly errors. This can also be related to a trade-off between reward and policy complexity[44], where less complex and cheaper to encode policies will lead to higher rates of choice perseveration (i.e., repeat choices). Time pressure has also been shown to increase participants' preferences for a known payoff over an uncertain alternative in the domain of gains[45], although the inverse was true in the domain of losses. There are also similar findings from description-based gambles, where time pressure can increase risk aversion in the domain of gains[46], with field experiments also showing that time pressure decreases risk-taking in auctions[47].

These divergent results could be interpreted through the lens of early work on coping mechanisms people use when put to the limits of their cognitive abilities[48]. One mechanism is *acceleration*, where information is processed at a faster rate, generating more errors. This can lead to choosing options more frequently that would otherwise be ignored, consistent with increased random exploration. Recent work using drift diffusion models have supported this hypothesis by connecting random exploration to lowered evidence thresholds and increased drift rates[49]. Conversely, longer response times have been related to the ability to mentally simulate a greater number of future outcomes[50], producing more directed exploration but decreased random exploration[51]. Acceleration as a response to time pressure could thus produce a trade-off between different forms of exploration.

Another potential mechanism is *repetition*, where previous actions are repeated or recycled[44,52], since it may not always be cost effective to simulate any future outcomes at all. This can be related to value-free habits[53], where not all decisions justify the cognitive costs of using value expectations (both rewards and uncertainty) to select new actions. Whereas you might normally enjoy exploring new restaurants in a new city, limits on decision time, such as an imminent departure at the airport, might motivate you to default to a previously visited restaurant, instead of weighing and selecting a new option.

Goals and Scope

We present a rich experimental setting where we use a within-subject design manipulating the presence or absence of time pressure to gain insights into the cognitive processes underlying exploration. We use multiple four-armed bandit tasks, where across four payoff conditions (within-subject), we independently manipulate reward expectations and uncertainty across different options (Fig. 1). This allows us to dissociate value-directed and uncertainty-directed choices, where compared to previous studies with two-armed bandit tasks[19,24], the richer set of options makes efficient exploration more relevant and observable over more trials. Given less decision time, participants can be expected to have less access to costly computations, leading to less value-maximizing choices and more random exploration. Simultaneously, time pressure may limit the capacity for reasoning about the uncertainty of each option, thus leading to less uncertainty directed exploration.

As predicted, time pressure made participants less sensitive to reward values (more random exploration)

and less likely to select options with high relative uncertainty (less directed exploration). We then estimated three hierarchical Bayesian models to understand how expectations of reward and subjective uncertainty influenced choices, reaction times (RTs), and evidence accumulation, which reaffirmed our behavioral analyses, with additional insights into the decision-making and evidence accumulation process.

Time pressure diminished uncertainty-directed exploration through several mechanisms: i) reducing the selection of uncertain options during early trials, ii) encouraging more aggressive exploitation of known options in later trials, iii) and heightening the tendency to repeat previous choices.

Our analysis of the RT data revealed how this shift in exploration is related to the computational costs of different exploration strategies. High reward expectations corresponded to faster choices, while high uncertainty (both relative and total) were associated with slower choices. Under time pressure, participants selected highly rewarding options even faster, but slowed down less when selecting highly uncertain options—independent of having faster choices in general. These changes in RT can be linked to the evidence accumulation process. While time pressure did not change how reward expectations influenced evidence accumulation (faster choices were due to lower decision-thresholds), it reduced the influence of relative uncertainty on evidence accumulation.

These results indicate that time pressure selectively impacts how uncertainty is integrated into decisions. Put under time pressure, people are less influenced by uncertainty, less value-directed, and more likely to repeat previous choices. This is a simpler strategy and comes at lower costs, representing a resource-rational adaptation to time pressure. These results enrich our understanding of human exploration strategies under changing task demands, providing insights into the cognitive costs of reasoning about and acting on uncertainty.

Results

We conducted an experiment to study how time pressure influences exploration behavior (Fig. 1). Our "Time Bandit" experiment employed repeated four-armed bandit tasks, where we independently manipulated expected reward and uncertainty across four payoff conditions (Fig. 1c; Table 1). This allowed us to disentangle how relative differences in reward expectations and uncertainty influence choices, and how time pressure modulates this influence, in a single within-subject design (see Methods).

Behavioral Analyses

Figure 2a depicts the average learning curves over participants, where one can see an increase in performance over trials for all payoff conditions (except Equal Means where no learning was possible). Overall, participants performed worse under time pressure, acquiring lower average rewards (paired t-test: $t(98) = -3.1$, $p = .002$, $d = 0.3$, $BF = 10$). This reduction in performance also holds when controlling for individual participant variability and simultaneously modelling the influence of payoff conditions in a Bayesian mixed effects regression (see Fig. S1).

Entropy and Repeat Choices

Consistent with the hypothesis that time pressure reduces exploration, participant choices were less diverse under limited time, as measured by the Shannon entropy[54] of the choices in each round ($t(98) = -4.1$, $p < .001$, $d = 0.4$, $BF > 100$; Fig. 2b). Lower entropy corresponds to less diversity of choices, where the minimum entropy strategy would be to always choose the same option. In contrast, the maximum entropy strategy would be to choose each option an equal number of times.

This difference in choice diversity can be partially attributed to the increased number of repeat choices that participants made under limited time ($t(98) = 6.2$, $p < .001$, $d = 0.6$, $BF > 100$; Fig. 2c). Looking more closely at repeat behavior, we find that in general, the probability of a repeated choice increased as a

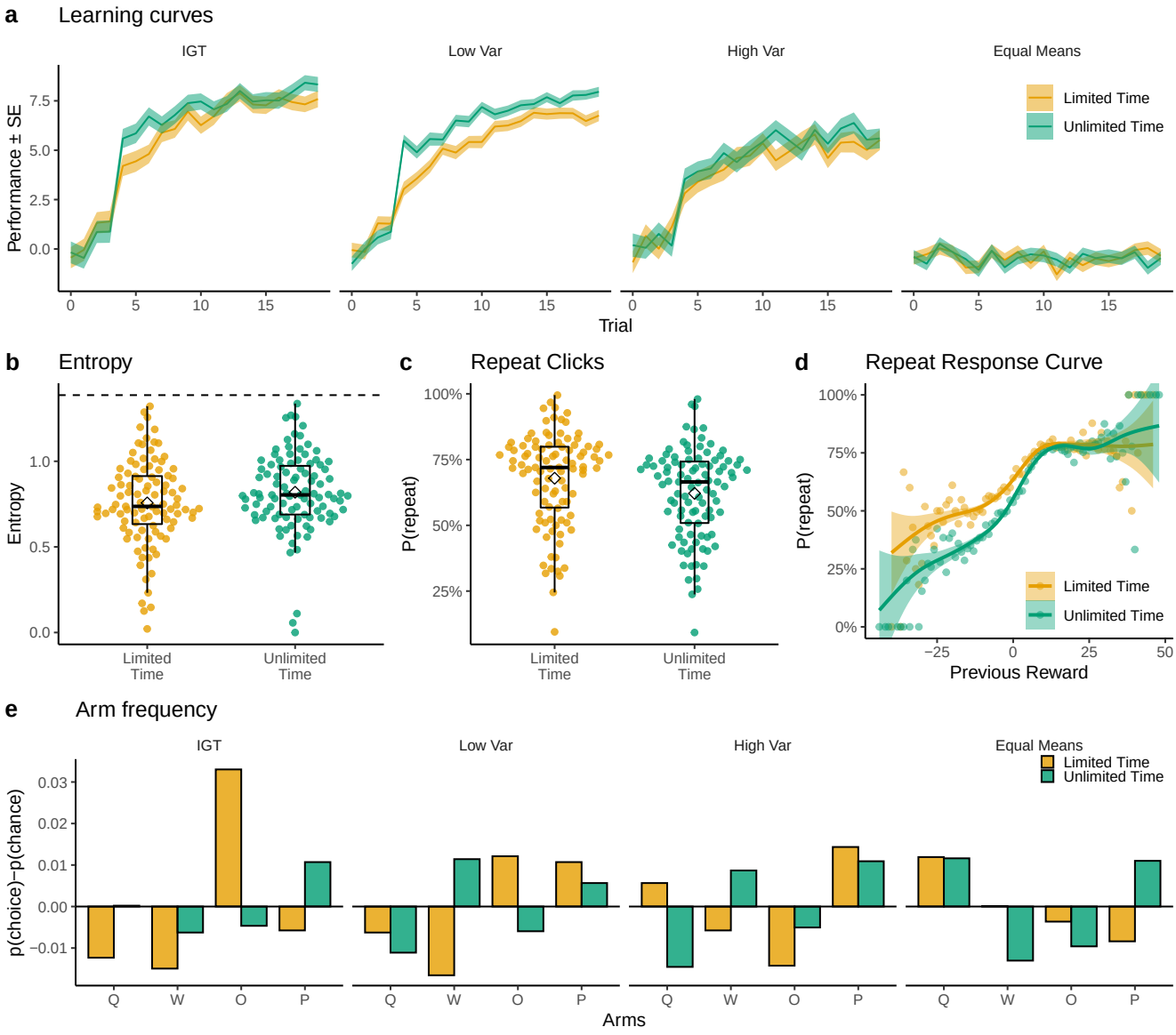

Figure 2. Payoff conditions and behavioral results. **a**) Learning curves of average participant performance (using unshifted rewards) over trials by payoff condition. Ribbons indicate standard error of the mean. **b**) The entropy of choices in each round, where higher entropy corresponds to more diverse choices and the dotted line indicates random chance (i.e., playing each arm with uniform probability). Each dot represents a participant, and overlaid are Tukey boxplots with the diamond indicating the group mean. **c**) The proportion of repeat clicks across time conditions, where each dot is a single participant, with overlaid Tukey boxplots and the diamond indicating the group mean. **d**) Repeat choices as a function of the previous (unshifted) reward value. Each dot is the aggregate mean, and lines represent a locally smoothed Generalized Additive Model regression estimate, with the ribbon indicating the 95% confidence interval. **e**) Aggregate choice proportions (normalized for chance) for each option, mapped to the canonical ordering shown in Fig. 1c). See Fig. S1 for a Bayesian mixed effects regression of the behavioral results, and see Fig. S2 and Fig. S3 for additional raw behavioral data.

function of the value of the reward obtained in the previous trial (average correlation: $\hat{r} = .56$; one-sample t-test comparing z-transformed coefficients against $\mu = 0$: $t(98) = 19.7$, $p < .001$, $d = 2.0$, $BF > 100$; Fig. 2d). Comparing the two time conditions, we find a steeper response curve and stronger correlation between previous reward value and repeat probability in unlimited time (paired t-test: $t(97) = -6.0$, $p < .001$, $d = 0.6$, $BF > 100$; one participant excluded due to undefined correlation). These results also hold under a Bayesian mixed effects regression (see Fig. S1). Thus, a source of increased repeat choices under limited time (and thus lower choice diversity) is due to participants being more likely to repeat options with lower-valued outcomes in the limited time condition.

Choice Patterns

To get a better sense of the choice patterns, we plot the aggregate choice proportions in Figure 2e using the canonical mapping of reward distributions to the $[Q, W, O, P]$ keys shown in Fig. 1c (randomly mapped in the experiment). These bars indicate the aggregate choice frequency of each option relative to chance, where bars above zero indicate the option was chosen more frequently, and bars below zero indicate the option was chosen less frequently. The difference between the orange and green bars illustrates the differences in choice behavior as a function of time pressure. To provide statistical support for choice differences, we use Bayesian mixed-effects logistic regression to model how time pressure influenced the probability of choosing a given option. We focus on two informative cases.

In the IGT condition (named for mimicking the structure of the so-called Iowa Gambling Task[55]), there were two high reward and two low reward options, with each pair having either a low or high variance. We focused on the two high reward options (indicated as 'O' and 'P' in Fig. 2e), and modeled whether time pressure influenced the likelihood of choosing the riskier high variance option 'P' over the safer low variance option 'O', as a simple test of how decision time can influence the role of relative uncertainty. We found that overall, participants chose the high variance option ('P') more frequently in unlimited time (Odds Ratio: OR $= 1.11$, 95% HPD: $[.80, 1.53]$; Table S2), although the estimates overlapped with chance (OR $= 1$). However, there was also an interaction with round number, where the difference between time conditions widened over successive rounds. Participants in the unlimited time condition increased their likelihood of selecting the high variance option over rounds (OR $= 1.39$, 95% HPD: $[1.23, 1.57]$). This effect tended towards the opposite direction for limited time rounds, where participants selected the high variance option less frequently over rounds (OR $= 0.83$, 95% HPD: $[0.68, 1.02]$).

We find the clearest differences arising from the time-pressure manipulation in the Equal Means condition, where compared against all other options, participants were more likely to select the highest variance option ('P') in the unlimited time condition (OR $= 1.44$, 95% HPD: $[1.12, 1.86]$; Table S2). This illustrates a clear shift in preferences away from uncertain options when time pressure is introduced. Whereas participants tend to be risk-seeking and choose more uncertain options under unlimited time, they become more risk-averse and choose them less often under time pressure.

Interim Discussion

Altogether, we find behavioral evidence that time pressure reduced exploration. There were less diverse and more repeat choices, which ultimately resulted in lower reward outcomes. From these analyses, we find two important behavioral signatures of the underlying cognitive processes that produced this shift in exploration. First, time pressure reduced participants' sensitivity to reward values in repeating previous choices, making them more likely to repeat options with low reward (Fig. 2d). Second, participants were less likely to select options with higher relative uncertainty under time pressure (Fig. 2e).

In the next section, we employ model-based analyses, which use RL models to explicitly track expected reward and uncertainty estimates. We then use these estimates to model choice behavior, reaction times, and evidence accumulation (i.e., drift rate).

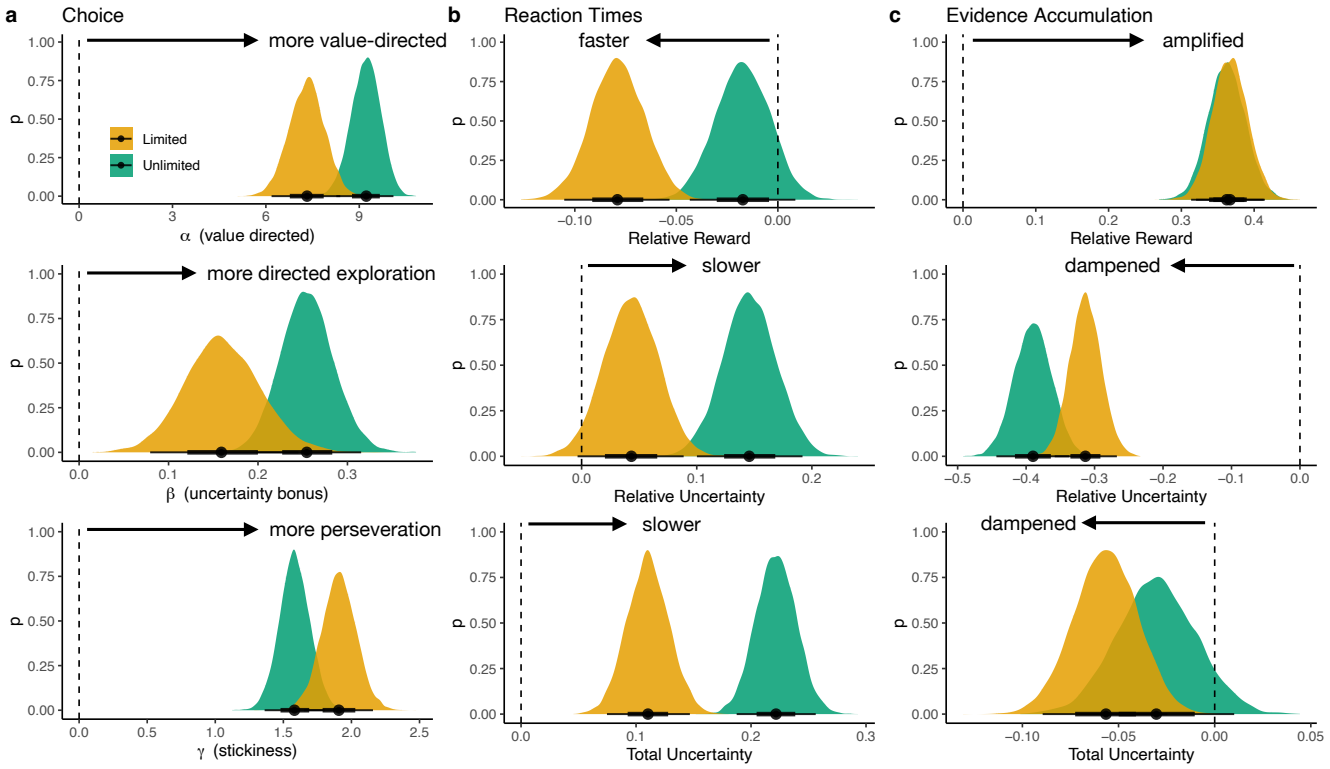

Figure 3. Posterior estimates of model-based analyses. a) Hierarchical softmax model. Expected rewards and uncertainties were regressed onto choice probability (Eqs 1-2). The top row shows the value-directed component (α), the middle row shows the uncertainty bonus (β), and the bottom row shows stickiness (γ). **b**) Hierarchical RT model. The influence of relative reward (top), relative uncertainty (middle), and total uncertainty (bottom) on RTs. **c**) LBA drift regression. Relative reward (top), relative uncertainty (middle), and total uncertainty (bottom) were used to predict the drift rate of an LBA. In all plots, the vertical dashed line indicates an effect of 0, while the black dot indicates the mean effect and confidence intervals show the 66% (thick) and 95% (thin) highest posterior density (HPD).

Model-Based Analyses

To model learning and decision making in our task, we use a *Bayesian mean tracker* (BMT) as an RL model for estimating expected rewards and associated uncertainties, which are then updated based on prediction errors (see Methods). The BMT is a form of Kalman filter, assuming time-invariant reward distributions, as was the case in our experiment. The BMT provides a Bayesian analogue of the classic Rescorla-Wagner[56] model of associative learning[57], and has described human behavior in a variety of multi-armed bandit and decision-making tasks[19,20,24,58–60].

We generated predictions from the BMT using participant choices and reward observations at each trial t to compute posterior distributions of the average reward of the options, and then using these as prior predictive distributions at trial $t+1$. These prior predictive distributions are all Normal, and we used the mean and standard deviations as measures of predicted reward and the associated uncertainty, respectively (see Fig. 4 and Fig. S4), to conduct three model-based analyses, predicting choices, reaction times, and evidence accumulation (Fig. 3).

Choices.

In our first analysis, we assessed how reward expectations and uncertainty estimates influenced the likelihood of an option being chosen on each trial. We applied hierarchical Bayesian inference to estimate

the parameters of a softmax policy (see Methods), under the assumption that a participant's choice on each trial is influenced by both the predicted mean and uncertainty of an option. Each participant's parameters are assumed to be jointly normally distributed and assumed to interact with time pressure. The probability of choosing option j on trial t is a softmax function of its decision value $Q_{j,t}$:

$$P(C_t = j) = \frac{\exp(Q_{j,t})}{\sum_{k=1}^{4} \exp(Q_{k,t})}. \quad (1)$$

The decision value $Q_{j,t}$ is a function of the prior predictive mean $m_{j,t}$ and uncertainty $\sqrt{v_{j,t}}$ (standard deviation) of each option according to the BMT, with an additional stickiness bonus for the most recently chosen option ($\delta_{j,t-1} = 1$ if option j was chosen on trial $t-1$; see Methods):

$$Q_{j,t} = \alpha(m_{j,t} + \beta\sqrt{v_{j,t}}) + \gamma\delta_{j,t-1}. \quad (2)$$

We computed hierarchical Bayesian estimates for the value-directed component α (factoring in both rewards and uncertainty), the uncertainty bonus β (governing the trade-off between exploitation and exploration), and the stickiness bonus γ, including interactions with the time pressure manipulation (limited vs. unlimited). Larger α estimates indicate more value-directed choices, whereas lower α suggest more random choices, which are not explainable by reward expectations or uncertainty estimates (i.e., random exploration). More positive β estimates indicate a higher level of uncertainty-directed exploration. Higher estimates of γ indicate more perseveration in choice behavior, with more frequent repetitions of previous choices. Figure 3a shows the posterior estimates of the model (see Fig. S6-S7 for comparison to alternative models).

We find less *value-directed* choice behavior under time pressure ($\alpha_{\text{Unlimited}} - \alpha_{\text{Limited}} = 1.90$, 95% HPD: $[1.24, 2.56]$), with positive estimates in both conditions ($\alpha_{\text{Unlimited}} = 9.21$, 95% HPD: $[8.31, 10.1]$; $\alpha_{\text{Limited}} = 7.31$, 95% HPD: $[6.23, 8.44]$). This pattern can be seen in the BMT predictions (Fig. 4a), where chosen options had both higher relative reward expectations and relative uncertainty in unlimited time. By definition, the inverse of the value-directed component defines the level of random exploration, with the interpretation that participants' choices were less predictable and more random when given less time to deliberate (limited time). This may seem at odds with the behavioral results showing reduced entropy under time pressure, but the lack of correlation between α and choice entropy under time limitations (see Fig. S5b) suggest that participants were consistently choosing non-value maximizing options. In contrast, α estimates were correlated with higher average rewards in both conditions (see Fig. S5a).

Time pressure also reduced *uncertainty-directed exploration* ($\beta_{\text{Unlimited}} - \beta_{\text{Limited}} = 0.09$, 95% HPD: $[0.04, 0.15]$), with positive estimates in both conditions ($\beta_{\text{Unlimited}} = .26$, 95% HPD: $[.20, .32]$; $\beta_{\text{Limited}} = .16$, 95% HPD: $[.08, .24]$). Figure 4b provides additional clarity about this result. Participants with unlimited time experienced an early uptick in selecting relatively uncertain options around trial 3, suggestive of an "exploration phase". Afterwards, there was a gradual shift towards exploitation, indicated by the monotonic decay of the relative uncertainty of chosen options, indicating an increasing preference for relatively less uncertain options. Under time pressure, there is a similar trend, yet the early exploration phase has almost vanished (the relative uncertainty of the chosen option on trial 3 is indistinguishable from 0: $t(98) = 0.9$, $p = .387$, $d = 0.1$, $BF = .16$) and later trials are associated with more strongly negative relative uncertainty. Thus, a reduced exploration bonus under time pressure appears to be a combination of less exploration in early trials, and more aggressive exploitation in later trials, which is also apparent in the higher levels of total uncertainty during limited time rounds (Fig. 4c). This reduction in directed exploration may also be related to the lower overall performance under time pressure, since higher β estimates in the limited time condition were associated with higher rewards (see Fig. S5a).

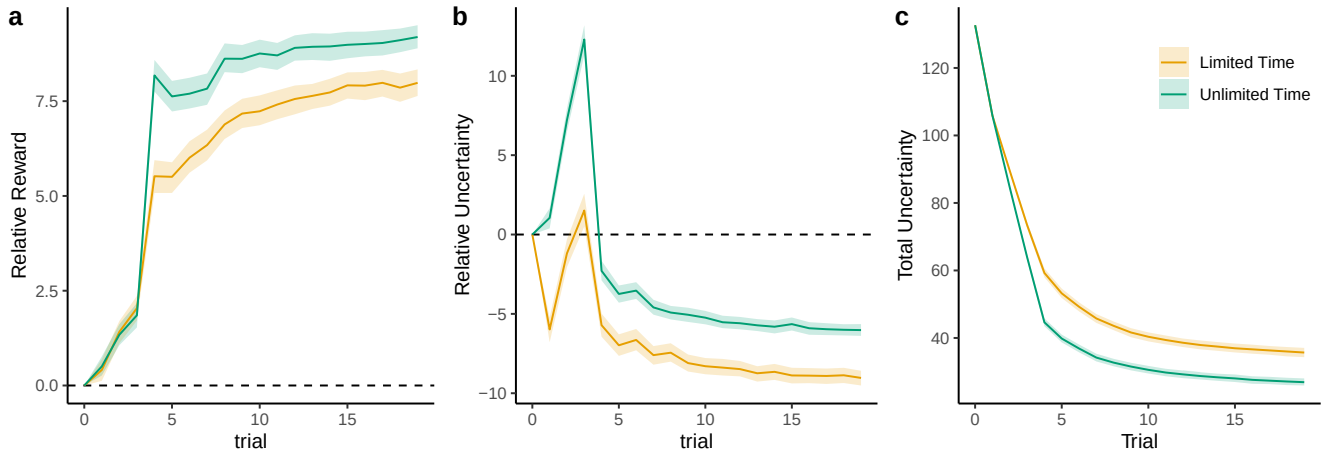

Figure 4. BMT predictions about the chosen option simulated for all participants (see Fig S4 for more detailed plots, separated by payoff condition). Lines indicate group means, with ribbons showing the 95% CI. **a**) Relative reward shows the difference between the posterior mean of the chosen option and the average posterior mean of the unchosen options. Relative reward is always valued positively (dashed line indicates 0). **b**) Relative uncertainty shows the difference between the posterior uncertainty (stdev) of the chosen option and the average posterior uncertainty of the unchosen options. The early upticks indicates uncertainty-directed exploration (substantially less in limited time), followed by exploitation as this value decays below zero (dashed line). **c**) Total uncertainty (average stdev) decays monotonically, with a faster decline in unlimited time due to more uncertainty directed exploration.

In addition to these changes in exploration, time pressure increased the *stickiness* of choices ($\gamma_{\text{Unlimited}} - \gamma_{\text{Limited}} = -0.32$, 95% HPD: $[-0.46, -0.19]$), with positive estimates in both conditions ($\gamma_{\text{Unlimited}} = 1.58$, 95% HPD: $[1.36, 1.80]$; $\gamma_{\text{Limited}} = 1.91$, 95% HPD: $[1.64, 2015]$). This increase in choice perseveration is consistent with the reduced entropy and higher repeat probabilities found in the behavioral data, but estimates of γ were unrelated to average reward (see Fig. S5).

Overall, time pressure reduced the value-directedness of choices, reduced uncertainty-directed exploration, and increased the stickiness of choices. We now turn to modeling reaction times (RTs) to better understand how reward expectations and uncertainty influenced the speed of decisions.

Reaction Time.

Our second analysis looked at how RTs (see Fig S3 for raw RT analysis) were influenced by expectations of rewards and estimated uncertainties. We first computed the relative reward and relative uncertainties of the BMT predictions using the difference between the chosen option and the average of the unchosen options on each trial. Thus, positive values indicate that the expected reward or uncertainty were larger than the mean of the unchosen options. We also computed total uncertainty, based on the sum of uncertainty estimates across the four options on any given trial. We then regressed relative mean, relative uncertainty, total uncertainty, and round number onto log-transformed RTs in a Bayesian mixed effects regression (see Methods).

The resulting posterior parameter estimates (Fig. 3b) indicate higher relative reward expectations produced faster choices under limited time ($b_{\text{Limited}} = -.08$, 95% HPD: $[-.11, -.05]$), but with a weaker effect under unlimited time ($b_{\text{Unlimited}} = -.02$, 95% HPD: $[-.04, .01]$) that overlapped with zero. In contrast, both relative and total uncertainty slowed down choices (relative uncertainty: $b_{\text{Unlimited}} = .15$, 95% HPD: $[.10, .19]$; total uncertainty: $b_{\text{Unlimited}} = .22$, 95% HPD: $[.19, .26]$), with the latter having a larger effect. In both cases, this uncertainty-related slowdown was reliably less pronounced when

placed under time pressure (relative uncertainty: $b_{\text{unlimited}} - b_{\text{limited}} = -.10$, 95% HPD: $[-.14, -.06]$; total uncertainty: $b_{\text{unlimited}} - b_{\text{limited}} = -.11$, 95% HPD: $[-.14, -.08]$). Thus, higher reward expectations made people faster, whereas uncertainty (both relative and total) slowed them down. Both effects were less pronounced under time pressure.

There was also a notable interaction between predictors (see Fig. S8 for the full model and Figs. S9-S10 for interaction plots). While high relative reward expectations generally sped up choices, this pattern was inverted when high rewards were also accompanied by high relative uncertainty, with participants slowing down instead of speeding up ($b = .04$, 95% HPD: $[.001, .08]$; no difference between time conditions). Thus, certainty about high rewards produced rapid decisions, whereas uncertainty about high rewards produced slower choices.

Overall, more exploitative choices (with higher relative reward expectations) were faster, while more explorative choices (with both higher relative uncertainty or higher total uncertainty) were slower. This differs from previous findings using two-armed bandits[19], in which higher relative uncertainty was related to faster decisions. Here, we find that uncertainty is not just a bonus that adds to the decision signal, making choices easier and faster, but rather, grappling with uncertainty takes time.

Evidence Accumulation.

In our third analysis, we used a Linear Ballistic Accumulator[61] (LBA) to model choices and reaction times simultaneously (see Methods). This model assumes that choices are the result of an evidence accumulation process, where evidence for each option accumulates continuously over time. Whichever option first exceeds the decision threshold is chosen. The interplay between the drift rate and the threshold captures how participants trade response speed for accuracy, with higher thresholds requiring more evidence and producing more value maximizing choices, yet slower responses. Thus, we can use the LBA to separate out how time pressure impacts evidence accumulation in terms of the rate of evidence accumulation and the amount of evidence collected.

Participants had lower relative evidence thresholds k under time pressure ($t(98) = -5.2$, $p < .001$, $d = 0.5$, $BF > 100$), consistent with the need to arrive at decisions more quickly. However in addition to merely reducing the evidence threshold, participants also had higher mean drift rates under limited time compared to unlimited time ($t(98) = 7.1$, $p < .001$, $d = 0.7$, $BF > 100$), suggesting an accelerated rate of evidence accumulation. The maximum pairwise difference between drift rates was also larger under limited time ($t(98) = 6.0$, $p < .001$, $d = 0.6$, $BF > 100$), suggesting larger separation between different options. Additionally, participants had shorter non-decision times τ ($t(98) = -4.6$, $p < .001$, $d = 0.5$, $BF > 100$) and less maximum starting evidence A ($t(98) = -7.8$, $p < .001$, $d = 0.8$, $BF > 100$), when placed under time pressure. All parameters where strongly correlated across limited and unlimited time conditions (Kendall rank correlations; all $r_\tau > .40$; $BF > 100$). Thus, our LBA results confirm the intuition that participants reached faster decisions at lower evidence thresholds when time limitations were imposed (see Fig. S11 for parameter estimates and Fig. S12 comparing parameters to behavior), but they also accumulated evidence faster and with larger separation between options.

In a final step, we sought to better understand the evidence accumulation process and how it changes with time pressure. Thus, we regressed the BMT predictions of relative expected reward, relative uncertainty, and total uncertainty for each option onto its estimated drift rate using a Bayesian mixed effects regression. Note that the LBA parameters are estimated on each round, thus the BMT predictions are averaged over trials, but nevertheless capture differences in the trajectory of learning and the independent manipulations of expected rewards and uncertainty in the four payoff conditions.

The result of this analysis (Fig. 3c) revealed that higher relative reward expectations amplified evidence accumulation equally for limited and unlimited time ($b = .36$, 95% HPD: $[.31, .41]$; no interaction

with time pressure: $b_{\text{Unlimited - Limited}} = -.005$, 95% HPD: $[-.061, .051]$). Thus, options with higher relative reward expectations were more likely to be chosen and with faster decision times. Conversely, relative uncertainty (specific to each option) had a negative effect on drift rate, thus dampening evidence accumulation ($b_{\text{Unlimited}} = -.39$, 95% HPD: $[-.44, -.34]$), with a reliably smaller effect under time pressure ($b_{\text{Limited}} = -.31$, 95% HPD: $[-.36, -.27]$; $b_{\text{Unlimited - Limited}} = -.08$, 95% HPD: $[-.14, -.01]$). Lastly, total uncertainty (computed across all options) also dampened evidence accumulation in limited time rounds ($b_{\text{Limited}} = -.06$, 95% HPD: $[-.09, -.03]$), but did not produce a reliable effect in unlimited time ($b_{\text{Unlimited}} = -.03$, 95% HPD: $[-.07, .01]$). Thus, rewards increased evidence accumulation, while uncertainty (in general) slowed down evidence accumulation.

The main interaction between predictors (see Fig. S13 for the full model and Figs. S14-S15 for interaction plots), was that the effect of total uncertainty could be inverted depending on relative reward (no interaction with time pressure: $b = -.08$, 95% HPD: $[-.14, -.02]$; Fig. S14g) and relative uncertainty ($b_{\text{Limited}} = -.15$, 95% HPD: $[-.18, -.12]$; $b_{\text{Unlimited}} = -.09$, 95% HPD: $[-.13, -.06]$; Fig. S14h). Total uncertainty amplified evidence accumulation when the stakes were low (low relative rewards or low relative uncertainty), but dampened evidence accumulation instead when the stakes were high (high relative rewards or relative uncertainty). Since total uncertainty is the same across all options, amplified evidence accumulation under low stakes corresponds to faster, more random choices, consistent with little benefit from increased deliberation in these settings. Conversely, dampened evidence accumulation under high stakes corresponds to slower, and more reward- or uncertainty-directed choices.

Overall, we find that reward-modulated increases in evidence accumulation were unaffected by time pressure. However, uncertainty-driven decreases in evidence accumulation were less pronounced under time pressure, with drift rates less influenced by uncertain options. We also found an influence of high total uncertainty, which was modulated by expectations of rewards and relative uncertainty. When more was at stake, total uncertainty dampened drift rates and produced slower decisions. But when relative differences in reward expectations were minor, higher total uncertainty amplified drift and produced faster decisions.

Discussion

How is exploration and decision-making constrained by cognitive limitations imposed through time pressure? We investigated this question using several variants of a four-armed bandit task, designed to independently manipulate differences in reward expectations and uncertainty. We then used a time pressure manipulation to either give participants unlimited decision time or to limit decision time to less than 400 ms for each choice. Both payoff and time pressure manipulations were conducted within-subjects, allowing us to use hierarchical modeling to achieve a high level of detail into the interplay between learning strategies and cognitive limitations imposed by time pressure.

Our behavioral results show that time pressure led participants to earn fewer rewards, made them less sensitive to reward values in their repeat choice behavior, and less likely to select options associated with higher uncertainty. We then used RL models to analyze how reward expectations and uncertainty affected choices, RTs, and the rate of evidence accumulation.

High reward expectations made participants more likely to select options, producing faster RTs for such exploitative choices, and amplifying the rate of evidence accumulation. Adding time pressure reduced the value-directedness of choices, but increased their tendency to speed up when choosing options with high relative reward expectations (i.e., exploitation), and made them more likely to repeat previous choices.

In contrast, while uncertainty also made participants more likely to select options, choices with higher relative and total uncertainty were associated with slower choices and reduced evidence accumulation

rates. Adding time pressure reduced uncertainty-directed exploration in choice behavior, thus reducing the influence of uncertainty on RTs and the rate of evidence accumulation (relative uncertainty only). This is consistent with the notion that uncertainty takes time to process and deploy strategically. Without the necessary time to grapple with uncertainty, participants shifted to exploiting known options and repeating previous choices, rather than integrating the value of exploring uncertain options.

Similar reductions in directed exploration have also been observed when participants were placed under working memory load[62]. The resulting behavior may thus be seen as a resource-rational[4,5] adaptation to externally imposed limitations on cognitive resources, consistent with other findings showing that people are sensitive to the cost-benefit tradeoffs of different learning strategies[63,64]. Indeed, the interactions of our LBA model (Fig. S14g-h) suggest that people are sensitive to the cost-benefit trade-off of increased deliberation, producing faster more random decisions when the stakes are low, but slowing down and deliberating longer when the stakes are high. Future research should examine the underlying mechanisms of the arbitration between strategies and the neural locus of cognitive control.

Limitations and extensions

One limitation is that we only account for how time pressure influences exploration strategies, but not for changes in learning. Time pressure might not only change which computations we engage in when deciding how to explore or exploit, but it might also influence the richness of the representations we form during learning or the extent to which these representations are updated in response to new information. Indeed, previous work in economics has shown a reduced efficacy of training[65]. However, our use of Bayesian RL in modeling choices and RTs may not be able to differentiate between these hypotheses, although similar models in related tasks have been used to predict directly elicited participant judgments about reward expectations and confidence[66–70]. Future studies may consider modeling not only choices and RTs, but also participant judgments about future outcomes in a similar time pressure manipulation.

Our current results also only examined uncertainty about reward expectations. However, there exist several alternative measures of uncertainty such as confidence[71,72], perceptual uncertainty[73,74], and computational uncertainty induced by cognitive load[75], all of which could influence exploration behavior in different ways. Thus, we expect future studies to increasingly focus on disentangling different sources of uncertainty and their effects on the exploration-exploitation dilemma.

Additionally, while our four-armed bandit task was designed to provide a richer choice set beyond two options, magnifying the difference between directed and random exploration, it still pales in comparison to the complexity of many real world problems. Since participants may be more likely to engage in directed exploration in highly complex or highly structured domains[21,22,70], an important future direction will be to understand how environmental structure modulates changes in learning as a function of cognitive limitations.

Lastly, we have also only looked at multi-armed bandits in which participants only gain positive rewards or earn nothing when exceeding the time limit. We did not, however, probe how exploration behavior changes in the domain of losses[76,77]. Since the distribution of rewards can affect participants' learning[78] and losses have been shown to produce risk-seeking under time pressure[45,46], studying this domain will be a crucial next step.

Conclusions

We studied the interplay of human exploration strategies and cognitive limitations imposed by time pressure, showing that participants are sensitive to the costs and benefits of different computations. Under time pressure, participants were less value-directed, less uncertainty-seeking, and more likely to repeat previous choices. These behavioral changes are linked to the cognitive costs of reasoning about rewards

and uncertainty, where in general, exploitative choices (i.e., high reward expectations) were faster, while exploratory choices (i.e, high relative or total uncertainty) were slower. Taken together, our results suggest that people display a resource-rational sensitivity to the cost-benefits of different exploration strategies under externally imposed limitations on cognitive resources.

Methods

Participants and Design.

We recruited 99 participants (36 female, aged between 21 and 69 years; M=34.82; SD=10.1) on Amazon Mechanical Turk (requiring 95% approval rate and 100 previously approved HITs). Participants were paid \$3.00 for taking part in the experiment and a performance contingent bonus of up to \$4.00 (calculated based on the performance of one randomly selected round). Participants spent 13.0 ± 5.6 minutes on the task and earned $\$5.87 \pm \$0.91$ in total. The study was approved by the Ethics Committee of the Max Planck Institute for Human Development.

We used a 2×4 within-subject design to examine how the presence or absence of time pressure and the payoff structure of the task (see Fig. 1c and Tab. 1) influenced choices and reaction times. In total, the experiment consisted of 40 rounds with 20 trials each. In each round, a condition was sampled (without replacement) from a pre-randomized list, such that each combination of time pressure and payoff structure was repeated five times, with a total of 100 trials in each.

Materials and Procedure.

Participants were required to complete three comprehension questions and two practice rounds (one with unlimited time and one with limited time) consisting of 5 trials each before starting the experiment. Each of the 40 rounds was presented as a four-armed bandit task, where the four options were randomly mapped to the $[Q, W, O, P]$ keys on the keyboard (Fig. 1a). Selecting an option by pressing the corresponding key yielded a reward sampled from a normal distribution, where the mean and variance was defined by the round's payoff structure (Fig. 1c and Tab. 1). Participants completed 20 trials in each round and were told to acquire as many points as possible.

Before starting a round, participants were informed whether it was an unlimited or a limited time round. In unlimited time rounds, participants could spend as much time as they needed to reach a decision, upon which they were given feedback about the obtained reward (displayed for 400 ms) before continuing to the next trial (Fig. 1b). In limited time rounds, participants were instructed to decide as fast as possible. If a decision took longer than 400 ms, they forfeited the reward they would have earned (presented to them as a crossed-out number with an additional sad smiley; Fig. 1b). We used the same inter-trial period of 400 ms to display feedback about obtained rewards in both limited and unlimited time rounds.

We applied a random shifting of rewards across rounds (i.e., different minimum and maximum reward) to prevent participants from immediately recognizing when they had chosen the optimal option. For each round, we sampled a value from a uniform distribution $\mathcal{U}(30, 60)$, which was then added to the rewards. Together with random shifting, we also truncated rewards such that they were always larger than zero. In order to convey intuitions about the random shift of rewards, payoffs were presented using a different fictional currency in each round (e.g., ß, Þ, ϑ), such that the absolute value was unknown, but higher were always better.

At the end of each round, participants were given feedback about their performance in terms of the bonus they would gain (in USD) if this was the round selected for determining the bonus. The bonus was calculated as a percentage of the total possible performance, raised to the power of 4 to accentuate differences in the upper range of performance: $\text{Bonus} = \left(\frac{\text{total reward gained}}{\text{mean reward of best option} \times 20 \text{ trials}}\right)^4 \times \$4.00$

Table 1. Payoff Conditions. Means shown are unshifted. In the experiment, a random value between 30 and 60 was added to all rewards of all options, and actual rewards were always positive.

Payoff Conds	Means (μ)	Variances (σ^2)
IGT	$[-10,-10,10,10]$	$[10,100,10,100]$
Low Var	$[-10,-\frac{1}{3},\frac{1}{3},10]$	$[10,10,10,10]$
High Var	$[-10,-\frac{1}{3},\frac{1}{3},10]$	$[100,100,100,100]$
Equal Means	$[0,0,0,0]$	$[10,40,70,100]$

Payoff conditions

We used four different payoff conditions as a within-participant manipulation (Table 1 and Fig. 1c). Each payoff condition specified the mean μ_j and variance σ_j^2 of the reward distribution $R_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for each option j. Each distribution was randomly mapped to one of the four $[Q,W,O,P]$ keys of the keyboard in each round. The Iowa Gambling Task (IGT) is a classic design that has been related to a variety of clinical and neurological factors affecting decision-making[55,79]. We implemented a reward condition inspired by the IGT such that there are two high and two low reward options, with a low and high variance version of each. We also constructed two conditions with equally spaced means, but with either uniformly low variance or uniformly high variance. Lastly, the equal means condition had identical means and gradually increasing variance, such that we can observe the influence of uncertainty independent of mean reward.

Model-based analyses

Bayesian mean tracker

The Bayesian mean tracker (BMT) learns a posterior distribution over the mean reward μ_j for each option j. Rewards are assumed to be normally distributed with a known variance but unknown mean. The prior distribution of the mean is also a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p_t(\mu_j|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \tag{3}$$

where p_t is the posterior distribution at trial t and $\mathcal{D}_{t-1}$ denotes the observed rewards and choices up to and including trial t (for all options). For a given option j, the posterior mean $m_{j,t}$ and variance $v_{j,t}$ at trial t are only updated when it has been selected at trial t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_t - m_{j,t-1} \right] \tag{4}$$

$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t} \right] v_{j,t-1} \tag{5}$$

where $\delta_{j,t} = 1$ if option j is chosen on trial t, and 0 otherwise. Additionally, y_t is the observed reward at trial t, and $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\varepsilon^2} \tag{6}$$

where θ_ε^2, referred to as the error variance, is the variance of the rewards around the mean.

Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on prediction error, which is the difference between the observed reward y_t and the prior expectation $m_{j,t-1}$, multiplied by learning rate $G_{j,t} \in [0,1]$. At the same time, the estimated variance $v_{j,t}$ of the chosen option is reduced by a factor $1 - G_{j,t}$. The error variance (θ_ε^2) can be interpreted as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$. We set the prior mean

to $m_{j,0} = 0$ based on the (unshifted) expectation across payoff conditions, and the prior variance is set to $v_{j,0} = 55 * 20$, which is also the expectation across payoff conditions, scaled by a constant multiple of 20. We use unshifted reward values (i.e., before adding the shift $\sim \mathcal{U}(30,60)$ were observed by participants), with the means in each condition centered on 0. For our model-based analysis, the error variance θ_ε^2 was set to the true underlying variance of the chosen option.

Hierarchical Bayesian Regression models

Mixed effects regressions

All Bayesian mixed effects regression models used Hamiltonian Markov chain Monte Carlo (MCMC) with a No-U-Turn sampler[80] and were implemented using `brms`[81]. All models used generic, weakly informative priors $\sim \mathcal{N}(0,1)$ with the proposal acceptance probability set to .99. In all cases, participants were assigned a random intercept and all fixed effects also had corresponding random effects following recommendations to apply a maximal random-effects structure[82]. All models were estimated over four chains of 4000 iterations, with a burn-in period of 1000 samples.

Softmax choice model

The softmax choice model was estimated hierarchically using custom code written in STAN. Formally, we assume that the α- and β-coefficients (see Equation 2) for each participant are drawn independently from a normal distribution:

$$\alpha_i^{\text{limited}}, \alpha_i^{\text{unlimited}}, \beta_i^{\text{limited}}, \beta_i^{\text{unlimited}}, \gamma_i^{\text{limited}}, \gamma_i^{\text{unlimited}} \sim \mathcal{N}(\mu_0, \sigma_0^2). \tag{7}$$

For simplicity, we use α_i, β_i, and γ_i in Equation 2 to refer to $\theta_i = \theta_i^{\text{limited}} + \mathbb{1}\theta_i^{\text{unlimited}}$, where $\theta \in [\alpha, \beta, \gamma]$ and $\mathbb{1} = 1$ for unlimited time rounds, and 0 otherwise. We used Hamiltonian MCMC with a No-U-Turn sampler[80] to estimate the group-level mean μ_0 and variance over participants σ_0^2 for α, β, and γ, and their interaction with time pressure. We used the following priors on the group-level parameters:

$$\mu_0 \sim \mathcal{N}(0,1) \tag{8}$$

$$\sigma_0^2 \sim \mathcal{N}(0,1) \in (0,\infty) \tag{9}$$

The posterior mean and uncertainty estimates of the BMT were standardized between [0,1] before being entered into the regression. The model was estimated over four chains of 4000 iterations, with a burn-in period of 1000 samples, and with the proposal acceptance probability set to .99.

RTs

The RT regression used the same Bayesian mixed effects framework as above, with log-transformed RTs as the dependent variable. 1 ms was added to each RT to avoid $\log(0)$, with the raw RTs truncated at a maximum of 5000 ms. Both dependent and independent variables were standardized to a mean of 0 and unit variance.

LBA

Formally, the LBA assumes that, after an initial period of non-decision time τ, evidence for option j accumulates linearly at a rate of v_j, starting from an initial evidence level $p_j \sim \mathcal{U}(0,A)$. Evidence accumulates for each option j until a threshold b is reached. We follow the Bayesian implementation proposed by Ref[83] and assume that the priors for the drift rates stem from truncated normal distributions

$$v_j \sim \mathcal{N}(2,1) \in (0,\infty). \tag{10}$$

Additionally, we assume a uniform prior on non-decision time

$$\tau \sim \mathcal{U}(0,1), \tag{11}$$

and a truncated normal prior on the maximum starting evidence

$$A \sim \mathcal{N}(0.5,1) \in (0,\infty). \tag{12}$$

Finally, we reparameterized the model by shifting b by k units away from A, and put a truncated normal distribution as the prior on the resulting relative threshold k:

$$k \sim \mathcal{N}(0.5,1) \in (0,\infty). \tag{13}$$

We estimated the LBA parameters (see Fig. S11) for each participant in every round separately using No-U-Turn Hamiltonian MCMC[80], with reaction times truncated at 5000 ms. The drift rate regression used the same Bayesian mixed effects framework as above, with both DVs and IVs standardized to a mean of 0 and unit variance.

Data and Code Availability

Code and data are publicly available at https://osf.io/v4dua/.

References

1. Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U. & Nieuwenhuis, S. The neural basis of the speed–accuracy tradeoff. *Trends neurosciences* **33**, 10–16 (2010).
2. Wickelgren, W. A. Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica* **41**, 67–85 (1977).
3. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. review* **117**, 1144 (2010).
4. Lieder, F. & Griffiths, T. L. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43** (2020).
5. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Curr. Opin. Behav. Sci.* **41**, 15–21 (2021).
6. Hertwig, R., Pleskac, T. J., Pachur, T. & Center for Adaptive Rationality. *Taming uncertainty* (Mit Press, 2019).
7. Mehlhorn, K. *et al.* Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decis.* **2**, 191 (2015).
8. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).
9. Cohen, J. D., McClure, S. M. & Yu, A. J. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philos. Transactions Royal Soc. B: Biol. Sci.* **362**, 933–942 (2007).
10. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018), second edn.
11. Lusena, C., Goldsmith, J. & Mundhenk, M. Nonapproximability results for partially observable markov decision processes. *J. Artif. Intell. Res.* **14**, 83–103 (2001).

12. Reverdy, P. B., Srivastava, V. & Leonard, N. E. Modeling human decision making in generalized gaussian multiarmed bandits. *Proc. IEEE* **102**, 544–571 (2014).

13. Gittins, J. C. & Jones, D. M. A dynamic allocation index for the discounted multiarmed bandit problem. *Biom.* **66**, 561–565 (1979).

14. Gittins, J. C. Bandit processes and dynamic allocation indices. *J. Royal Stat. Soc. Ser. B (Methodological)* 148–177 (1979).

15. Lai, T. L. & Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv.* **6**, 4–22 (1985).

16. Pleskac, T. J. Learning models in decision making. In Keren, G. & Wu, G. (eds.) *The Wiley Blackwell Handbook of Judgment and Decision Making*, vol. 2, 629–657 (Wiley Blackwell, 2015).

17. Wilson, R. C., Bonawitz, E., Costa, V. D. & Ebitz, R. B. Balancing exploration and exploitation with information and randomization. *Curr. Opin. Behav. Sci.* **38**, 49–56 (2021).

18. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 155–164 (2014).

19. Gershman, S. J. Uncertainty and exploration. *Decis.* (2018).

20. Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).

21. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* **2**, 915—-924 (2018). DOI 10.1038/s41562-018-0467-4.

22. Schulz, E. *et al.* Structured, uncertainty-driven exploration in real-world consumer choice. *Proc. Natl. Acad. Sci.* 201821028 (2019).

23. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Networks* **15**, 549–559 (2002).

24. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cogn.* **173**, 34–42 (2018).

25. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995* (2009).

26. Kahneman, D. & Frederick, S. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics biases: The psychology intuitive judgment* **49**, 81 (2002).

27. Ariely, D. & Zakay, D. A timely account of the role of duration in decision making. *Acta psychologica* **108**, 187–207 (2001).

28. Klein, G. Sources of error in naturalistic decision making tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 37, 368–371 (SAGE Publications Sage CA: Los Angeles, CA, 1993).

29. Donkin, C., Little, D. R. & Houpt, J. W. Assessing the speed-accuracy trade-off effect on the capacity of information processing. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1183 (2014).

30. Drugowitsch, J., DeAngelis, G. C., Angelaki, D. E. & Pouget, A. Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *Elife* **4**, e06678 (2015).

31. Bogacz, R., Hu, P. T., Holmes, P. J. & Cohen, J. D. Do humans produce the speed–accuracy trade-off that maximizes reward rate? *The Q. J. Exp. Psychol.* **63**, 863–891 (2010).

32. Olschewski, S., Rieskamp, J. & Scheibehenne, B. Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *J. Exp. Psychol. Gen.* **147**, 462 (2018).

33. Olschewski, S. & Rieskamp, J. Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *J. Behav. Decis. Mak.* (2021).

34. Madan, C. R., Spetch, M. L. & Ludvig, E. A. Rapid makes risky: Time pressure increases risk seeking in decisions from experience. *J. Cogn. Psychol.* **27**, 921–928 (2015).

35. Hayden, B. Y. & Platt, M. L. Temporal discounting predicts risk sensitivity in rhesus macaques. *Curr. Biol.* **17**, 49–53 (2007).

36. Hu, Y., Wang, D., Pang, K., Xu, G. & Guo, J. The effect of emotion and time pressure on risk decision-making. *J. Risk Res.* **18**, 637–650 (2015).

37. Huber, O. & Kunz, U. Time pressure in risky decision-making: effect on risk defusing. *Psychol. Sci.* **49**, 415 (2007).

38. Kocher, M. G., Pahlke, J. & Trautmann, S. T. Tempus fugit: time pressure in risky decisions. *Manag. Sci.* **59**, 2380–2391 (2013).

39. Maule, A. J., Hockey, G. R. J. & Bdzola, L. Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta psychologica* **104**, 283–301 (2000).

40. Young, D. L., Goodie, A. S., Hall, D. B. & Wu, E. Decision making under time pressure, modeled in a prospect theory framework. *Organ. behavior human decision processes* **118**, 179–188 (2012).

41. Gershman, S. J. & Bhui, R. Rationally inattentive intertemporal choice. *Nat. communications* **11**, 1–8 (2020).

42. Rieskamp, J. & Otto, P. E. Ssl: a theory of how people learn to select strategies. *J. Exp. Psychol. Gen.* **135**, 207 (2006).

43. Betsch, T., Haberstroh, S., Molter, B. & Glöckner, A. Oops, i did it again—relapse errors in routinized decision making. *Organ. behavior human decision processes* **93**, 62–74 (2004).

44. Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cogn.* **204**, 104394 (2020).

45. Busemeyer, J. R. Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 538 (1985).

46. Nursimulu, A. D. & Bossaerts, P. Risk and reward preferences under time pressure. *Rev. Finance* **18**, 999–1022 (2013).

47. El Haji, A., Krawczyk, M., Sylwestrzak, M. & Zawojska, E. Time pressure and risk taking in auctions: A field experiment. *J. behavioral experimental economics* **78**, 68–79 (2019).

48. Miller, J. G. Information input overload and psychopathology. *Am. journal psychiatry* **116**, 695–704 (1960).

49. Feng, S. F., Wang, S., Zarnescu, S. & Wilson, R. C. The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Sci. reports* **11**, 1–15 (2021).

50. Dasgupta, I., Schulz, E. & Gershman, S. J. Where do hypotheses come from? *Cogn. psychology* **96**, 1–25 (2017).

51. Wilson, R., Wang, S., Sadeghiyeh, H. & Cohen, J. D. Deep exploration as a unifying account of explore-exploit behavior. *PsyArXiv* (2020).

52. Thorndike, L. *Animal intelligence: Experimental studies* (1911).

53. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. review* **126**, 292 (2019).

54. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).

55. Bechara, A., Damasio, A. R., Damasio, H. & Anderson, S. W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cogn.* **50**, 7–15 (1994).

56. Rescorla, R. A. & Wagner, A. R. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class.* **2**, 64–99 (1972).

57. Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* **11**, e1004567 (2015).

58. Yu, A. J. & Dayan, P. Expected and unexpected uncertainty: ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems*, 173–180 (2003).

59. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Learning and decisions in contextual multi-armed bandit tasks. In *Thirty-Seventh Annual Conference of the Cognitive Science Society* (2015).

60. Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nat. neuroscience* **3**, 1218–1223 (2000).

61. Brown, S. D. & Heathcote, A. The simplest complete model of choice response time: Linear ballistic accumulation. *Cogn. Psychol.* **57**, 153–178 (2008).

62. Cogliati, I. D., Cleeremans, A. & Alexander, W. Should we control? the interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. experimental psychology. Gen.* (2019).

63. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. science* **28**, 1321–1333 (2017).

64. Aboody, R., Zhou, C. & Jara-Ettinger, J. In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Dev.* (2021).

65. Zakay, D. & Wooler, S. Time pressure, training and decision effectiveness. *Ergonomics* **27**, 273–284 (1984).

66. Boldt, A., Blundell, C. & De Martino, B. Confidence modulates exploration and exploitation in value-based learning. *bioRxiv* 236026 (2017).

67. Stojić, H., Schulz, E., P Analytis, P. & Speekenbrink, M. It's new, but is it good? how generalization and uncertainty guide the exploration of novel options. *J. Exp. Psychol. Gen.* **149**, 1878 (2020).

68. Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J. & Speekenbrink, M. Uncertainty in learning, choice, and visual fixation. *Proc. Natl. Acad. Sci.* **117**, 3291–3300 (2020).

69. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Comput. Biol.* **16**, 1–28 (2020). DOI 10.1371/journal.pcbi.1008149.

70. Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. *Comput. Brain & Behav.* **4**, 125–147 (2021). DOI 10.1007/s42113-020-00091-x.

71. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly bayesian analysis of confidence in perceptual decision-making. *PLoS computational biology* **11**, e1004519 (2015).

72. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. review* **117**, 864 (2010).

73. Bitzer, S., Park, H., Blankenburg, F. & Kiebel, S. J. Perceptual decision making: drift-diffusion model is equivalent to a bayesian model. *Front. human neuroscience* **8**, 102 (2014).

74. Bruckner, R., Heekeren, H. R. & Ostwald, D. Belief states and categorical-choice biases determine reward-based learning under perceptual uncertainty. *bioRxiv* (2020). DOI 10.1101/2020.09.18.303495.

75. Deck, C. & Jahedi, S. The effect of cognitive load on economic decision making: A survey and new experiments. *Eur. Econ. Rev.* **78**, 97–119 (2015).

76. Fontanesi, L., Palminteri, S. & Lebreton, M. Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cogn. Affect. & Behav. Neurosci.* **19**, 490–502 (2019).

77. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. communications* **6**, 1–14 (2015).

78. Gershman, S. J. Do learning rates adapt to the distribution of rewards? *Psychon. bulletin & review* **22**, 1320–1327 (2015).

79. Yechiam, E., Busemeyer, J. R., Stout, J. C. & Bechara, A. Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychol. Sci.* **16**, 973–978 (2005).

80. Hoffman, M. D. & Gelman, A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).

81. Bürkner, P.-C. Advanced bayesian multilevel modeling with the r package brms. *The R J.* **10**, 395–411 (2018).

82. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. memory language* **68**, 255–278 (2013).

83. Annis, J., Miller, B. J. & Palmeri, T. J. Bayesian inference with Stan: A tutorial on adding custom distributions. *Behav. Res. Methods* **49**, 863–886 (2017).

84. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. & Rev.* **16**, 225–237 (2009).

85. Jeffreys, H. *The Theory of Probability* (Oxford, UK: Oxford University Press, 1961).

86. Ly, A., Verhagen, J. & Wagenmakers, E.-J. Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *J. Math. Psychol.* **72**, 19–32 (2016).

87. van Doorn, J., Ly, A., Marsman, M. & Wagenmakers, E.-J. Bayesian inference for kendall's rank correlation coefficient. *The Am. Stat.* **72**, 303–308 (2018).

88. Zellner, A. & Siow, A. Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M., Lindley, D. V. & Smith, A. F. M. (eds.) *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, 585–603 (University of Valencia, 1980).

89. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Royal Soc. Lond. Ser. A. Math. Phys. Sci.* **186**, 453–461 (1946).

90. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default bayes factors for anova designs. *J. Math. Psychol.* **56**, 356–374 (2012).

91. Gelman, A. *et al. Bayesian data analysis* (CRC press, 2013).

Acknowledgments
We thank Rahul Bhui and Irene Cogliati Dezza for helpful feedback on earlier drafts of the manuscript, and Kimberly Gerbaulet for help with data collection. CMW is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

Author Contributions
All authors designed the experiment. CMW collected and analyzed the data, with contributions from ES. CMW drafted the initial manuscript with input from ES, TJP, and MS. All authors contributed to final manuscript.

Competing financial interests
The authors declare no competing financial interests.

Supporting information

Statistics

Comparisons.

Both frequentist and Bayesian statistics are reported throughout this paper. Frequentist tests are reported as Student's t-tests (specified as either paired or independent). Each of these tests are accompanied by a Bayes factors (BF) to quantify the relative evidence the data provide in favor of the alternative hypothesis (H_A) over the null (H_0). This is done using the default two-sided Bayesian t-test for either independent or dependent samples, where both use a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$, as suggested by Ref[84]. All statistical tests are non-directional as defined by a symmetric prior.

Correlations.

For testing linear correlations with Pearson's r, the Bayesian test is based on Jeffreys[85] test for linear correlation and assumes a shifted, scaled beta prior distribution $B(\frac{1}{k}, \frac{1}{k})$ for r, where the scale parameter is set to $k = \frac{1}{3}$[86]. Note that when performing group comparisons of correlations computed at the individual level, we report the mean correlation and the statistics of a single-sample t-test comparing the distribution of z-transformed correlation coefficients to $\mu = 0$.

For testing rank correlations with Kendall's tau, the Bayesian test is based on parametric yoking to define a prior over the test statistic[87], and performing Bayesian inference to arrive at a posterior distribution for r_τ. The Savage-Dickey density ratio test is used to produce an interpretable Bayes Factor.

ANOVA.

We use a two-way analysis of variance (ANOVA) to compare the means of $p \geq 2$ samples based on the F distribution. In general terms, we can define ANOVA as a linear model:

$$\mathbf{y} = \mu\mathbf{1} + \sigma\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{14}$$

where $\mathbf{y}$ is a vector of N observations, μ is the aggregate mean, $\mathbf{1}$ is a column vector of length N, σ is the scale factor, $\mathbf{X}$ is the $N \times p$ design matrix, $\boldsymbol{\theta}$ is a column vector of the standardized effect sizes, and $\boldsymbol{\varepsilon}$ is a column vector containing the i.i.d. errors where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

We assume independent g-priors[88] for each effect size $\theta_1 \sim \mathcal{N}(0, g_1\sigma^2), \cdots, \theta_p \sim \mathcal{N}(0, g_p\sigma^2)$, where each g-value is drawn from an inverse chi-square prior with a single degree of freedom $g_i \overset{\text{i.i.d}}{\sim}$ inverse-$\chi^2(1)$. For μ and σ^2 we assume a Jeffreys[89] prior. Following Ref[90], we compute the Bayes factor by integrating the likelihoods with respect to the prior on parameters, where Monte Carlo sampling was used to approximate the g-priors. The Bayes factor reported in the text can be interpreted as the log-odds of the model relative to an intercept-only null model.

Supplementary Behavioral results

Time Pressure and Payoff Manipulations

To analyze the influence of time pressure and payoff conditions on performance and choice behavior, we constructed a series of Bayesian mixed effects regression models. Specifically, we estimated how average rewards (Fig. 2a), the entropy of choices (Fig. 2b), the number of repeat choices (Fig. 2c), and the probability of making a repeat choice (Fig. 2e), were influenced by time pressure and payoff conditions, while also accounting for individual differences in the random effects structure (see Table S1).

We describe the influence of either time pressure or payoff conditions in terms of the estimated marginal means (Δ_{EMM}), which uses contrast analysis to describes the difference in the dependent variable marginalized over the other independent variables. For instance, examining how average rewards were

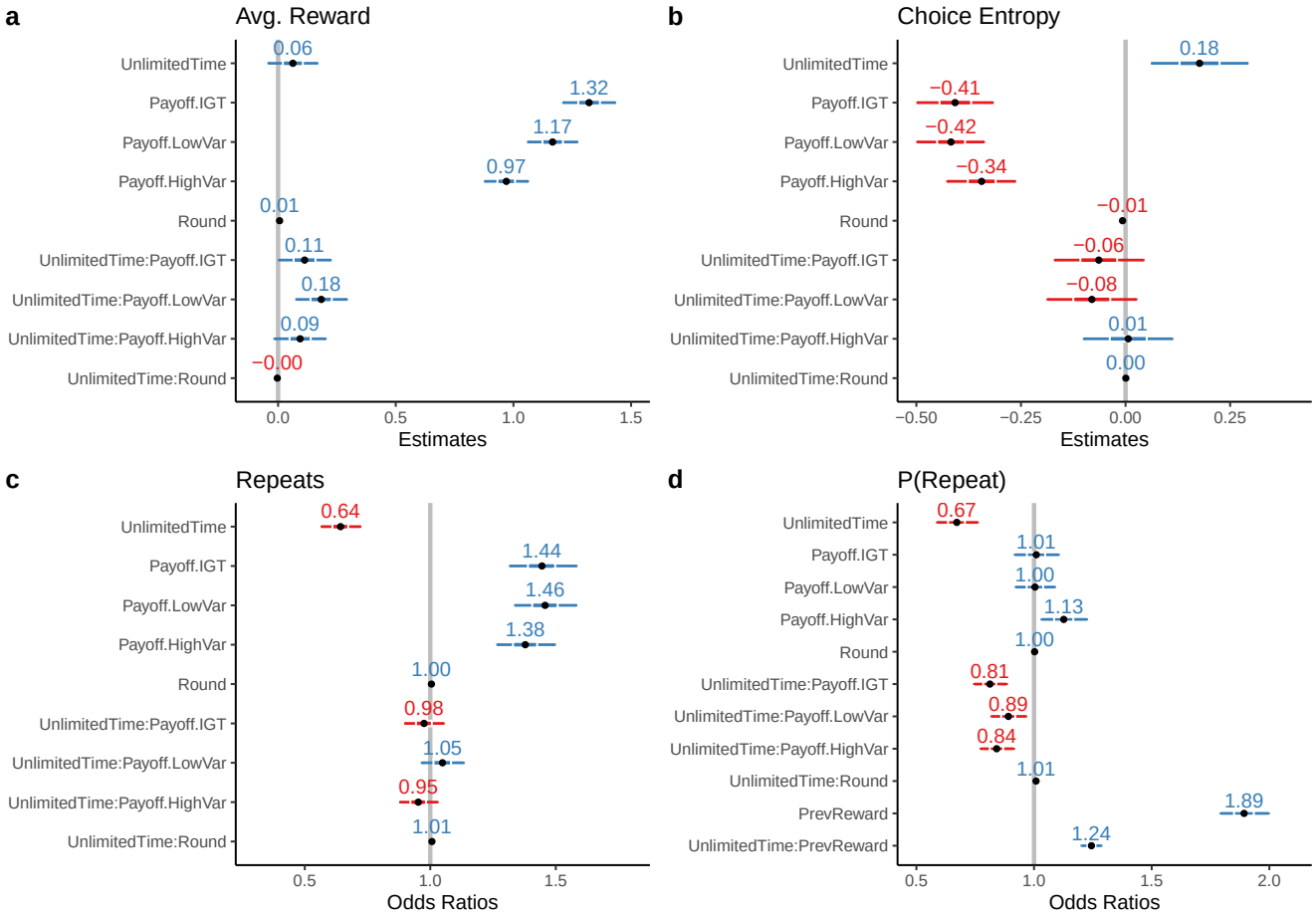

Figure S1. Regression coefficients. Visualization of regression coefficients, corresponding to the models in Table S1. The vertical grey line represents chance, and estimates above are in blue, while estimates below are in red (irrespective of significance). The inner horizontal line indicates the 50% HDI and the outer line indicates 89% HDI.

influenced by time pressure, marginalized over the four payoff conditions. The raw posterior estimates are provided in Table S1 and visualized in Figure S1. See Fig. S2 for the raw data, separated by condition.

Looking first at average reward, we find that the behavioral changes induced by time pressure played a reliable role in reducing rewards ($\Delta_{EMM} = -.19$, 95% Highest Density Interval (HDI): $[-0.29, -0.10]$). There was also substantial variation in average reward across payoff conditions. Participants performed better in the IGT-like condition than in the low variance condition ($\Delta_{EMM} = 0.12$, 95% HDI: $[0.04, 0.20]$). We see an even larger difference when comparing the Low Var and High Var conditions ($\Delta_{EMM} = .24$, 95% HPD: $[0.15, 0.33]$). Low Var and High Var had the same expected rewards for each option, yet participants performed substantially better with lower variance. Lastly, participants performed better in High Var than in the Equal Means condition ($\Delta_{EMM} = 1.02$,95% HDI: $[0.93, 1.11]$), which is intuitive since improvement is not possible if all arms have the same expected reward.

Next, we assessed the overall diversity of choices by calculating the Shannon entropy[54] of choices in each round. Participants made lower entropy choices under limited time ($\Delta_{EMM} = -0.12$, 95% HDI $= [-.23 - 0.01]$). This provides additional evidence for reduced exploration under time pressure, since we find a lower diversity of choices. We also find largely overlapping entropy levels among the different payoff conditions, but with Equal Means having the most diverse choices (compared against High Var: $\Delta_{EMM} = 0.34$, 95% HDI $= [0.27, 0.42]$). This suggests that in the face of indiscernible differences in

reward expectations, participants increased their exploration.

Additionally, we modeled the number of repeat choices in each round as a measure of the level of sequential dependency between choices. We used a Binomial regression, modeling the number of repeats as the result of 19 independent Bernoulli trials, since the first choice cannot be a repeat by definition. Participants made more repeat choices in the limited in the limited time condition (Odds Ration (OR): $\Delta_{EMM} = 1.40$, 95% HDI $= [1.22, 1.58]$). Thus, a reduction in decision time produced more sequentially dependent choices. We see relatively small variation in repeat choices across the payoff conditions. However, Low Var produced more repeats than High Var (OR: $\Delta_{EMM} = 1.34$, 95% HDI $= [1.23, 1.47]$), perhaps because participants were able to more quickly identify and exploit the highest rewarding arm with less variance in observed outcomes.

Lastly, we also included a variant of the repeat choice model, which included the (unshifted) value of the previous reward as an additional predictor. Here, we modeled the probability of each choice (after the first trial) being a repeat using logistic regression. We find the same influence of the experimental manipulations on repeat behavior as above (see Table S1), but also find an interaction between time pressure and previous reward value. Participants were more likely to make a repeat choice for higher rewards in unlimited time (OR= 1.24, 95% Highest Posterior Density (HPD) interval: $[1.19, 1.29]$). Put differently, time pressure reduced participants' sensitivity to reward value in their repeat behavior, as evidenced by the flatter response curve in Figure 2d.

Raw RTs.

Figure S3a shows the distribution of participant reaction times (RTs) split by time pressure and payoff conditions. Using a two-way within subject ANOVA, we found that participants (unsurprisingly) responded faster in limited time ($F(1,98) = 13.8$, $p < .001$, $\eta^2 = .016$, $BF > 100$), but with no differences across payoff conditions ($F(3,98) = 0.684$, $p = .562$, $\eta^2 = .002$, $BF = 0.005$). Additionally, we find that participants sped up over trials (average correlation: $\bar{r} = -.54$; one-sample t-test against zero using z-transformed correlation coefficients: $t(98) = -17.6$, $p < .001$, $d = 1.8$, $BF > 100$; Fig. S3b), with a strong speed up in unlimited time (paired t-test comparing z-transformed correlation coefficients: $t(98) = 4.5$, $p < .001$, $d = 0.5$, $BF > 100$). We see a similar speed-up over rounds (average correlation: $t(98) = -9.4$, $p < .001$, $d = 0.9$, $BF > 100$; Fig. S3c), which was also more pronounced under unlimited time ($t(98) = 4.4$, $p < .001$, $d = 0.5$, $BF > 100$).

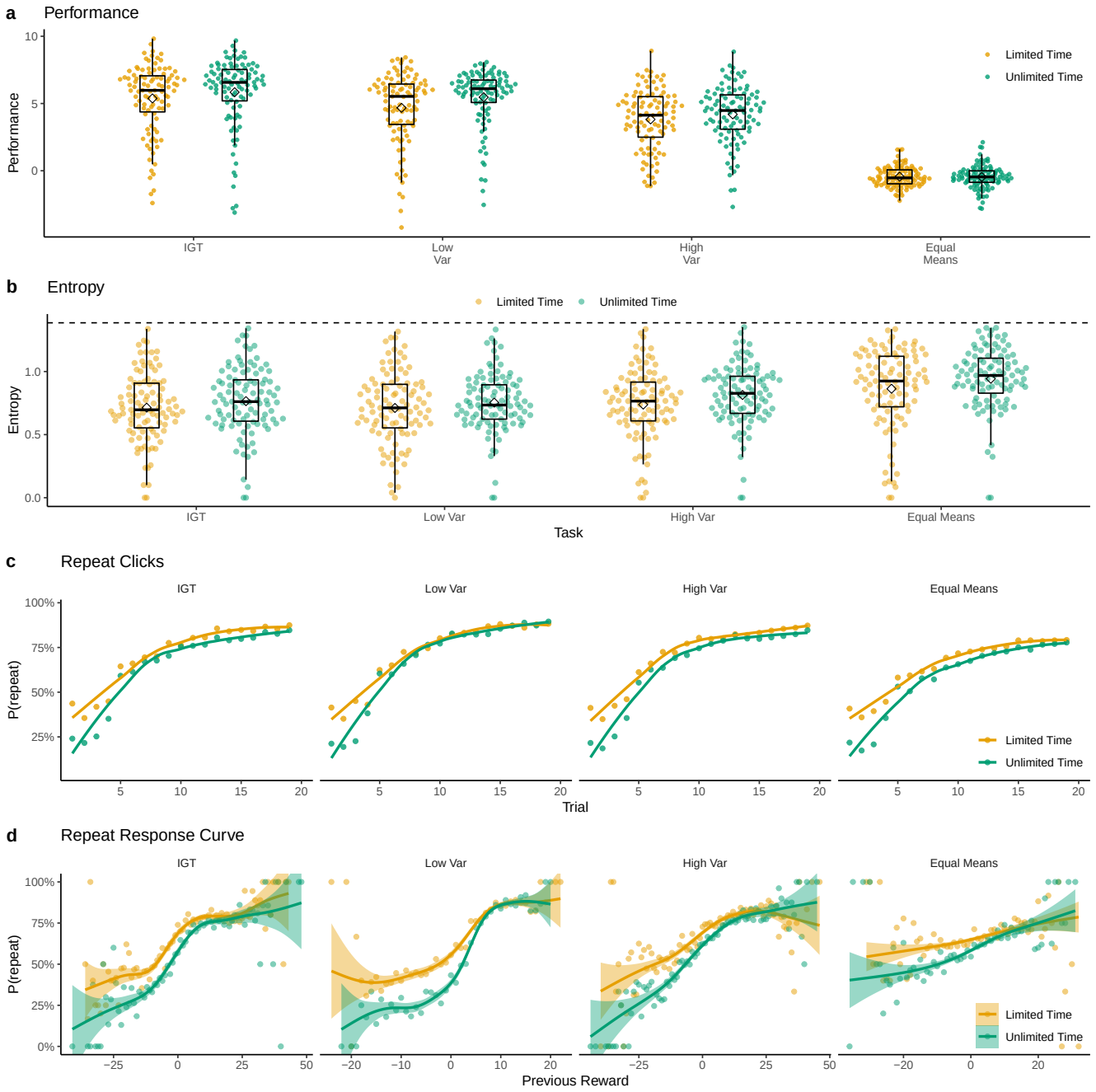

Figure S2. Raw Behavioral Data Split By Conditions. a) Average performance across payoff and time conditions. Each dot is a single participant, with overlaid Tukey boxplots and diamonds indicating the group means. **b**) Entropy of choices by payoff and time conditions, where the dashed line indicates a random baseline. **c**) Repeat clicks as a function of trial, where each dot is the group mean and the lines are a locally smoothed regression. **d**) Repeat clicks as a function of previous reward value. Ribbons indicate the 95% CI.

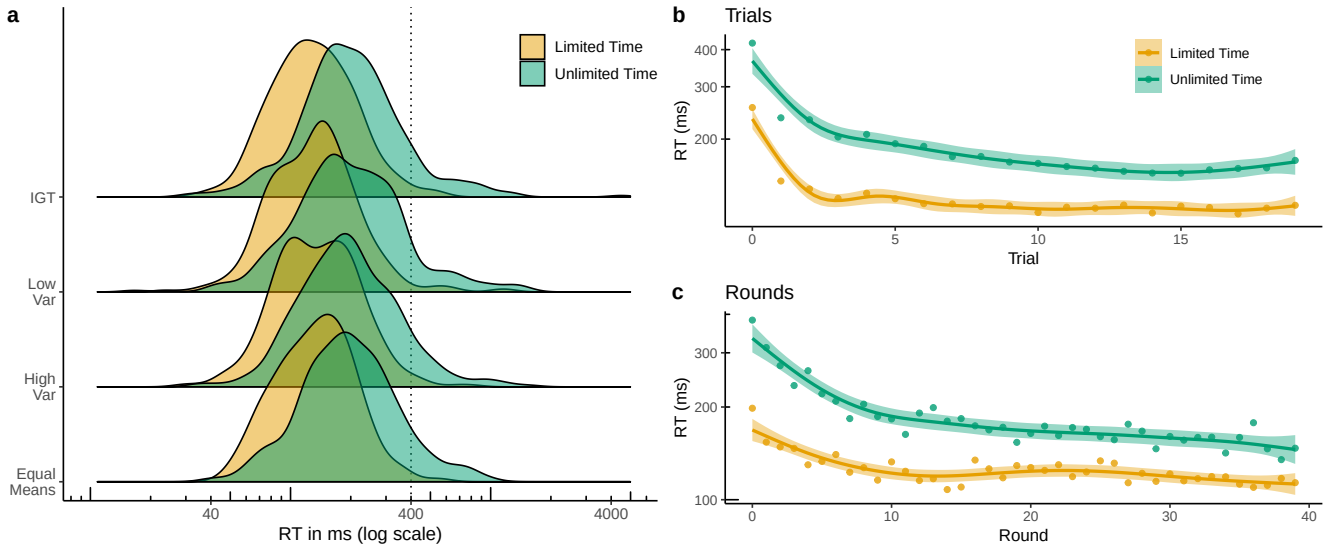

Figure S3. Reaction Times (RTs). All RTs are shown in milliseconds (ms) and on a log scale. Outliers greater than 5000ms are omitted from the plots, but not from the analyses. **a**) RT distributions separated by payoff and time conditions. The dashed line indicates the 400 ms limit for limited time choices. **b**) RTs as a function of trial. Each dot is the aggregate mean, with the lines and ribbons indicating the mean and 95% confidence intervals of a generalized additive regression. **c**) RTs as a function of Round.

Table S1. Bayesian Mixed Effects Regression: Experimental Manipulations

	Avg. Reward *Estimate*	Choice Entropy *Estimate*	Repeats *Odds Ratio*	P(Repeat) *Odds Ratio*
Intercept	-1.04	0.36	1.94	2.73
	[-1.13, -0.95]	[0.19, 0.52]	[1.55, 2.45]	[2.13, 3.47]
UnlimitedTime	0.06	0.18	0.64	0.67
	[-0.06, 0.19]	[0.04, 0.32]	[0.55, 0.74]	[0.57, 0.78]
Payoff.IGT	1.32	-0.41	1.44	1.01
	[1.19, 1.46]	[-0.52, -0.30]	[1.29, 1.61]	[0.90, 1.13]
Payoff.LowVar	1.17	-0.42	1.46	1.00
	[1.04, 1.29]	[-0.52, -0.32]	[1.31, 1.61]	[0.91, 1.11]
Payoff.HighVar	0.97	-0.34	1.38	1.12
	[0.86, 1.08]	[-0.44, -0.25]	[1.25, 1.52]	[1.01, 1.25]
Round	0.01	-0.01	1.00	1.00
	[0.00, 0.01]	[-0.01, -0.00]	[1.00, 1.01]	[1.00, 1.01]
UnlimitedTime:Payoff.IGT	0.11	-0.06	0.97	0.81
	[-0.2, 0.25]	[-0.19, 0.07]	[0.88, 1.07]	[0.73, 0.90]
UnlimitedTime:Payoff.LowVar	0.18	-0.08	1.05	0.89
	[0.05, 0.32]	[-0.21, 0.05]	[0.95, 1.15]	[0.81, 0.98]
UnlimitedTime:Payoff.HighVar	0.09	0.01	0.95	0.84
	[-0.04, 0.22]	[-0.12, 0.14]	[0.87, 1.04]	[0.76, 0.93]
UnlimitedTime:Round	-0.00	0.00	1.01	1.01
	[-0.01, 0.00]	[-0.00, 0.00]	[1.00, 1.01]	[1.01, 1.01]
PrevReward				1.89
				[1.77, 2.02]
UnlimitedTime:PrevReward				1.24
				[1.19, 1.29]
Random Effects				
σ^2	0.13	0.41	12.69	0.01
τ_{00}	0.88	0.59	4.47	0.21
ICC	0.13	0.41	0.72	0.04
$N_{Participant}$	99	99	99	99
Observations	3960	3960	3960	76240
Bayesian R^2	0.43	0.46	0.65	0.26

Note: Each model was defined as `DV ~ TimePressure * PayoffConditions * Round + (1 + TimePressure + PayoffConditions + Round | Participant)`, where `DV` is the dependent variable (columns), and `PayoffConditions` were defined using dummy coding, with the baseline being the Equal Means condition. We report the posterior mean and 95% highest posterior density (HPD) interval below in brackets. The 'Repeats' model is a Binomial regression based on 19 successive Bernoulli trials (since the first trial cannot be a repeat). The 'P(Repeat)' model is a logistic regression, with the previous reward value added as an additional predictor. Both the 'Repeats' and 'P(Repeat)' models are reported as Odds Ratios. σ^2 indicates the individual-level variance, τ_{00} indicates the variation between individual intercepts and the average intercept, and ICC is the intraclass correlation coefficient. Model coefficients are visualized in Figure S1.

Table S2. Bayesian Mixed Effects Logistic Regression: Choice Probability for Highest Variance Option

	IGT *Odds Ratio*	Equal Means *Odds Ratio*
Intercept	0.91	0.24
	[0.69, 1.20]	[0.18, 0.33]
UnlimitedTime	1.11	1.45
	[0.80, 1.53]	[1.11, 1.87]
Round	0.83	1.00
	[0.68, 1.02]	[0.99, 1.02]
UnlimitedTime:Round	1.39	0.99
	[1.23, 1.57]	[0.98, 0.99]
Random Effects		
σ^2	0.00	0.02
τ_{00}	0.25	0.17
ICC	0.00	0.00
$N_{Participant}$	99	99
Observations	10230	19800
Bayesian R^2	0.194	0.134

Note: Each model was defined as `DV` $\sim$ `TimePressure * Round + (1 + TimePressure + Round|Participant)`, where `DV` is the dependent dependent binary variable representing whether the highest variance option was chosen. In the IGT regression, we only consider choices where the two highest mean reward options were chosen ('O' and 'P'), where $DV = 1$ when 'P' was chosen, and zero otherwise. For the Equal Means condition, we include all choices. We report the posterior odds ratio and 95% highest posterior density (HPD) interval below in brackets. σ^2 indicates the individual-level variance, τ_{00} indicates the variation between individual intercepts and the average intercept, and ICC is the intraclass correlation coefficient.

Supplementary Model Results

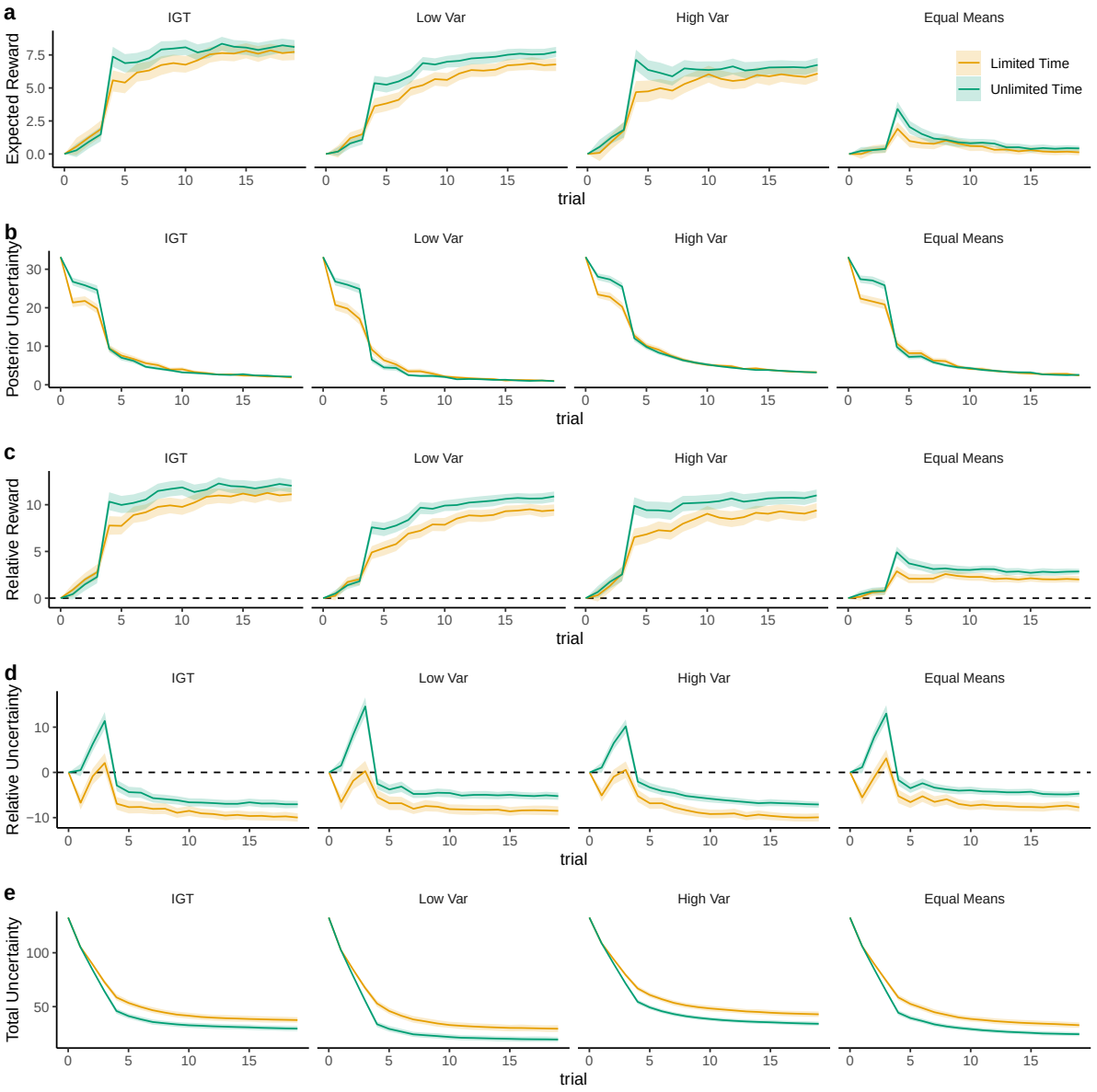

Figure S4. BMT predictions about the chosen option simulated for all participants. Lines indicate group means, with ribbons showing the 95% CI. **a**) Expected rewards (posterior mean) increase over successive trials, showing how the model tracks learning. The uptick in the Equal Means condition, followed by a decay back to zero indicates participants persevered after high reward observations stemming from the underlying variance, which then regressed back to the mean of 0. **b**) Posterior uncertainty (stdev) decays as participants exploit options with diminishing uncertainty. **c**) Relative reward shows the difference between the posterior mean of the chosen option and the average posterior mean of the unchosen options. Relative reward is always valued positively (dashed line indicates 0). **d**) Relative uncertainty shows the difference between the posterior uncertainty (stdev) of the chosen option and the average posterior uncertainty of the unchoen options. The early upticks indicates uncertainty directed exploration (substantially less in limited time), followed by exploitation as this value decays below zero (dashed line). **e**) Total uncertainty (stdev) decays monotonically, with a faster decline in unlimited time due to more uncertainty directed exploration.

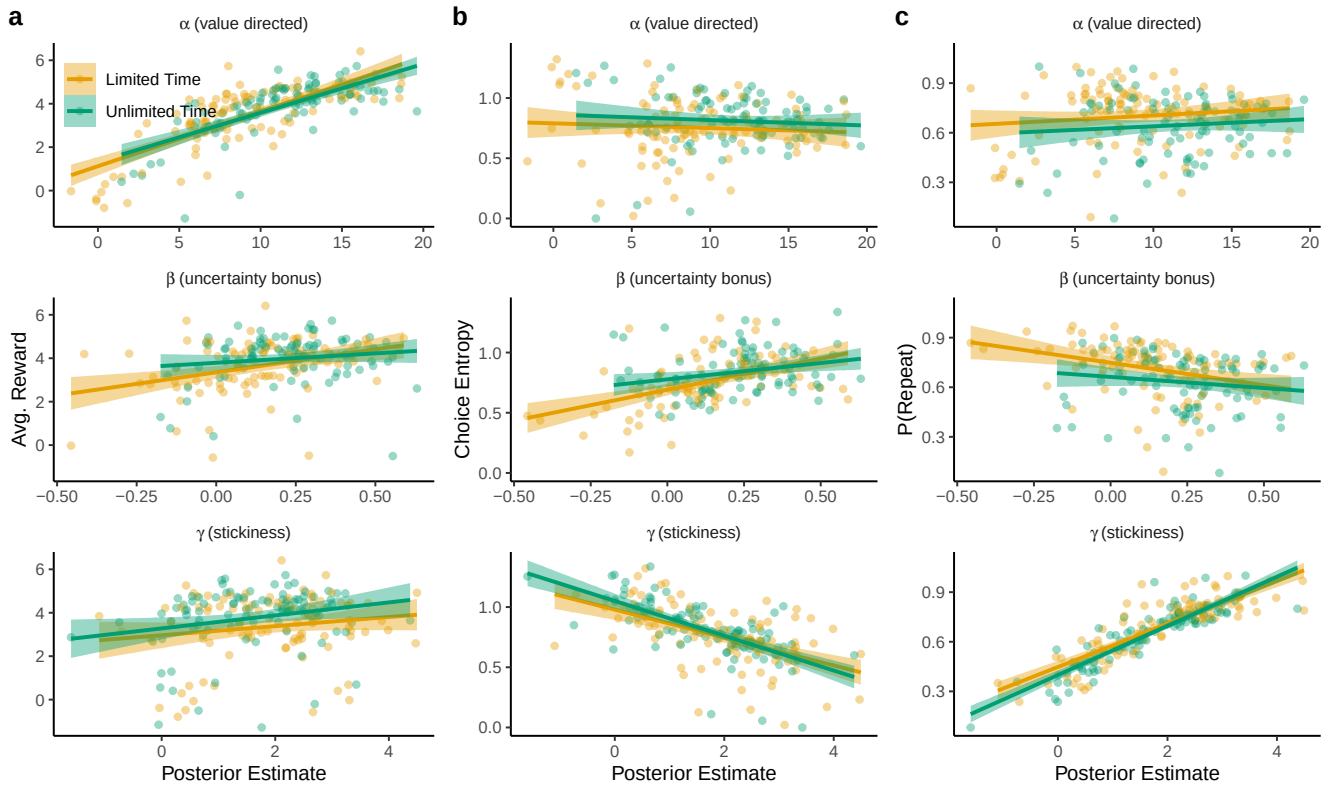

Figure S5. Comparison of Softmax parameters and behavior. Each dot shows the mean posterior for each participant in each time condition, while the lines and ribbons are a linear regression and 95% CI. **a**) Higher α estimates correspond to higher rewards in both conditions (unlimited time: $r_\tau = .55$, $p < .001$, $BF > 100$; limited time: $r_\tau = .55$, $p < .001$, $BF > 100$). However, we only find a reliable effect of β under time pressure ($r_\tau = .25$, $p < .001$, $BF > 100$), but not with unlimited time ($r_\tau = .13$, $p = .055$, $BF = .81$). This suggests that the lower overall performance under time pressure, may have a result of the reduction in uncertainty directed exploration (Fig. 3a). We find no relationship between stickiness and rewards (unlimited time: $r_\tau = .13$, $p = .052$, $BF = .85$; limited time: $r_\tau = .08$, $p = .265$, $BF = .24$). **b**) We find no correlation between α and choice entropy (unlimited time: $r_\tau = -.13$, $p = .053$, $BF = .84$; limited time: $r_\tau = -.01$, $p = .835$, $BF = .13$). However, higher β estimates generated higher entropy choices in both conditions (unlimited time: $r_\tau = .26$, $p < .001$, $BF > 100$; limited time: $r_\tau = .36$, $p < .001$, $BF > 100$), while higher γ were related to lower entropy (unlimited time: $r_\tau = -.53$, $p < .001$, $BF > 100$; limited time: $r_\tau = -.41$, $p < .001$, $BF > 100$). **c**) Similar to choice entropy, we find no relationship between α and the frequency of repeat choices (unlimited time: $r_\tau = .06$, $p = .384$, $BF = .19$; limited time: $r_\tau = -.04$, $p = .545$, $BF = .16$). However, higher β estimates were correlated with less repeat choices in limited time ($r_\tau = -.30$, $p < .001$, $BF > 100$), and more weakly correlated in unlimited time ($r_\tau = -.19$, $p = .006$, $BF = 5.4$). Stickiness γ was unsurprisingly correlated with more repeat choices in both conditions (unlimited time: $r_\tau = .73$, $p < .001$, $BF > 100$; limited time: $r_\tau = .65$, $p < .001$, $BF > 100$). In all plots, Tukey's fence has been applied to omit outliers for clearer visualizations, but all data are included in the statistical tests.

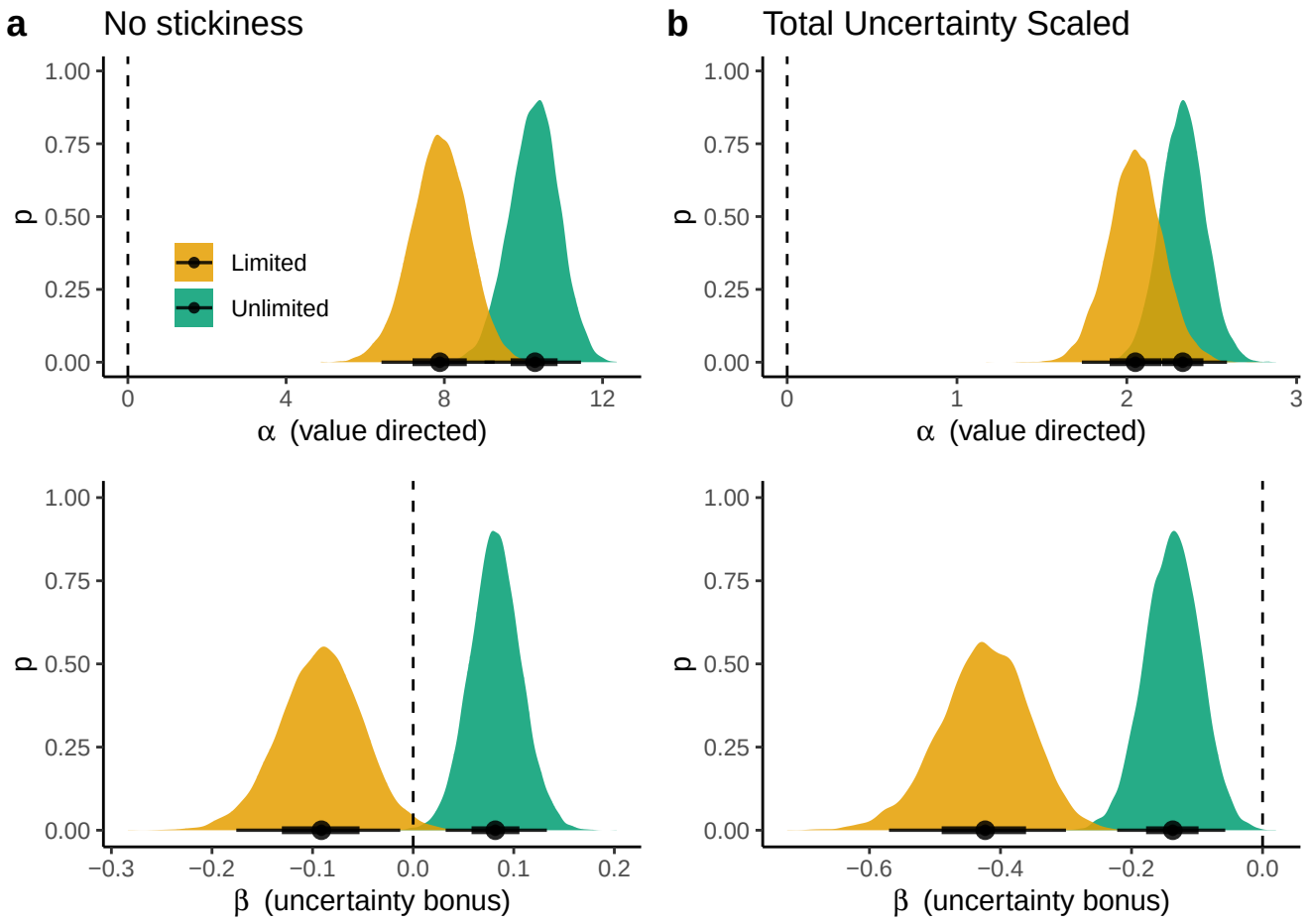

Figure S6. Alternative Softmax Model Posteriors. Posterior estimates for alternative formulations of the softmax model. **a**) Variant without stickiness, which yields negative uncertainty bonus β estimates for limited time. **b**) Variant, also without stickiness, but where the value-directed component was scaled by the total uncertainty (across options) as a method to regulate higher random exploration when the total uncertainty is high (following Ref[24]; see Fig. S7 for details). Here, we get negative uncertainty bonus β estimates for both conditions. Both models provide worse fits to the data (Fig. S7) compared to the sticky softmax model reported in Figure 3a .

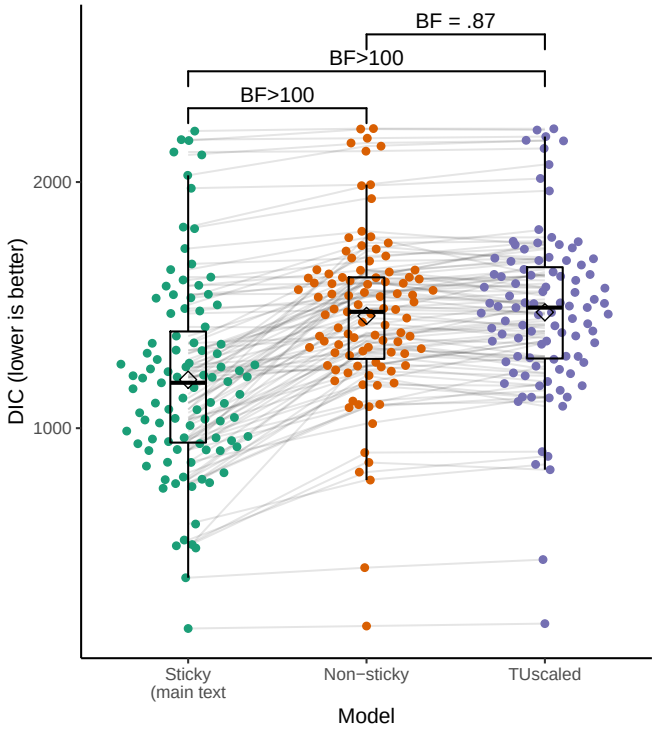

Figure S7. Model comparison. Comparing three variants of the hierarchical softmax choice model using Deviance Information Criterion[91] $DIC = -2\log\mathbb{E}_\theta p(y|\theta) + 2p_D$, where the effective number of parameters is defined as $p_D = \mathbb{V}_\theta\left(-2\log p\left(y|\theta\right)\right)$. The sticky model is reported in the main text, while the non-sticky model omits the γ term. Lastly, the total uncertainty scaled model (TUscaled) also omits the stickiness parameter, but rescales the value estimates going into the softmax function by the total uncertainty across all four options to account for changes in random exploration as a function of total uncertainty[24]: $Q_{j,t} = \frac{\alpha(m_{j,t}+\beta\sqrt{v_{j,t}})}{\sum_k \sqrt{v_{k,t}}}$. Each dot is a single participant (connected by lines across models),with overlaid Tukey boxplots and the diamond indicating the group mean. The significance tests are Bayes Factors (BF) corresponding to paired Bayesian t-tests. The sticky model beats the non-sticky model ($t(98) = -13.2$, $p < .001$, $d = 0.7$, $BF > 100$), the sticky model beats the TUscaled model ($t(98) = -15.6$, $p < .001$, $d = 0.7$, $BF > 100$), and there are no reliable differences between the non-sticky and TUscaled models ($t(98) = -2.1$, $p = .040$, $d = 0.0$, $BF = .87$).

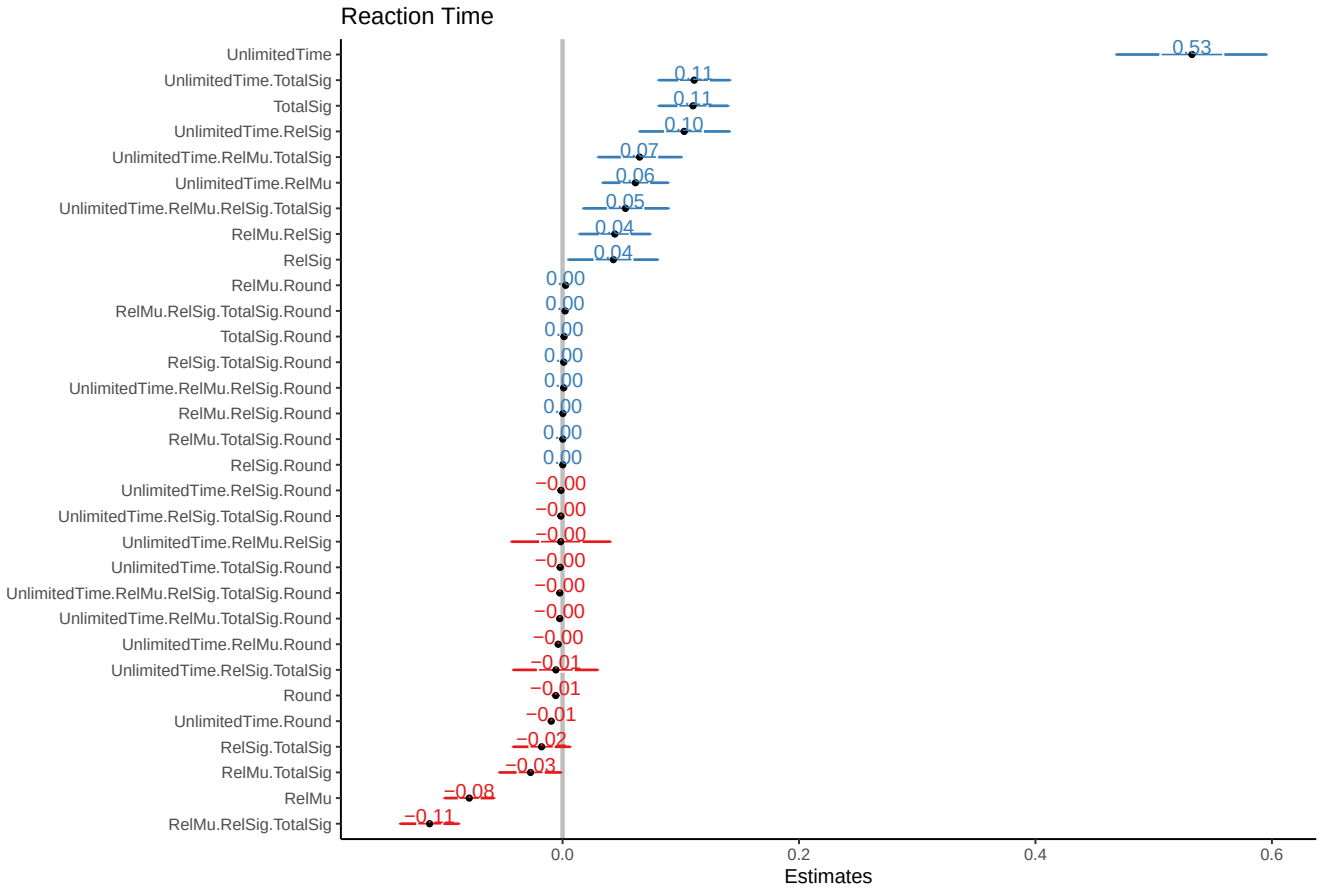

Figure S8. Coefficient plot of RT mixed effects regression. Posterior estimates of the Bayesian mixed effects regression predicting (log) RT. The mean posterior estimate is displayed numerically and indicated by the black dot, while the 95% HPD is illustrated by the length of the horizontal line. Coefficients are sorted by largest to smallest, with blue and red colors corresponding to estimates that are above or below 0, respectively, but do not indicate whether the difference is meaningful. See Figs. S9-S10 for interaction plots. RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.

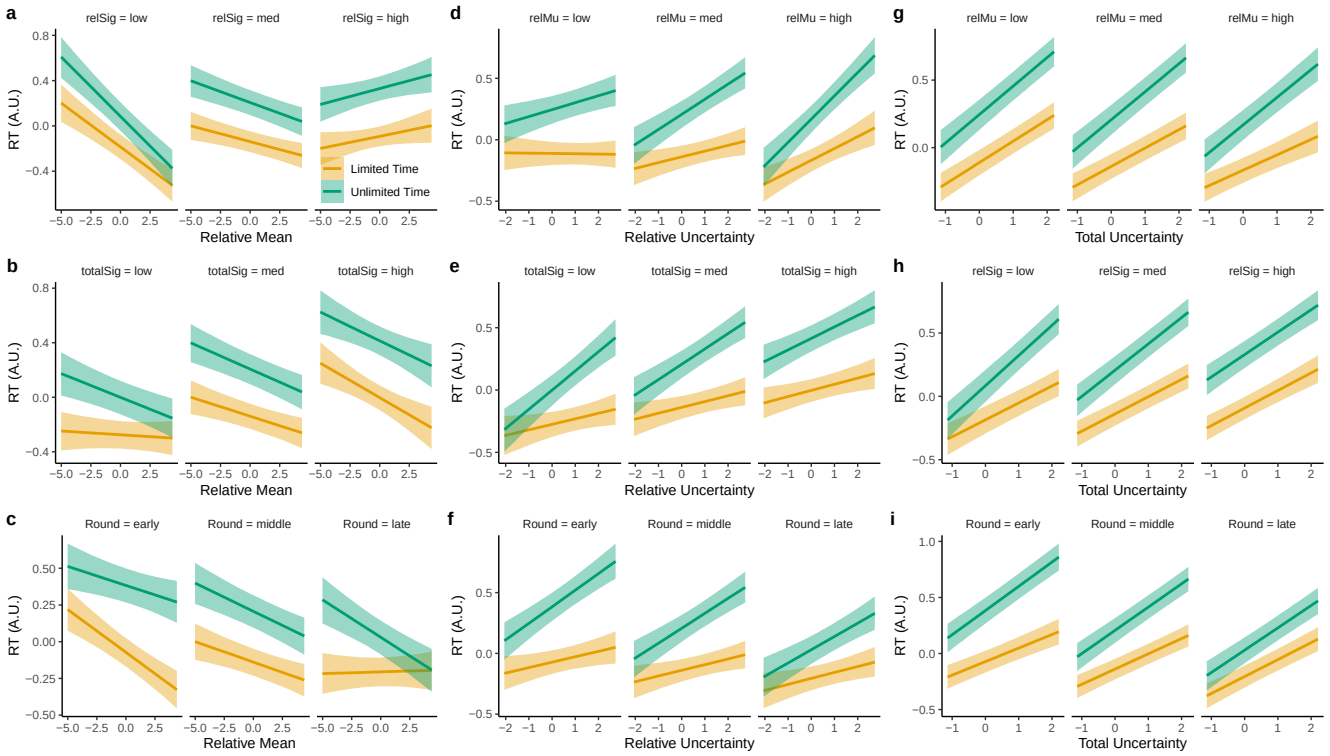

Figure S9. Marginal interaction plots for RT mixed effects regression. Marginal interactions of the RT Bayesian mixed effects regression illustrated in Fig. S8. Interactions are grouped in terms of relative means (**a-c**), relative uncertainty (**d-f**), and total uncertainty (**g-i**). Continuous variables are split into discrete $[low, med, high]$ levels, based on $[mean - sd, mean, mean + sd]$. RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.

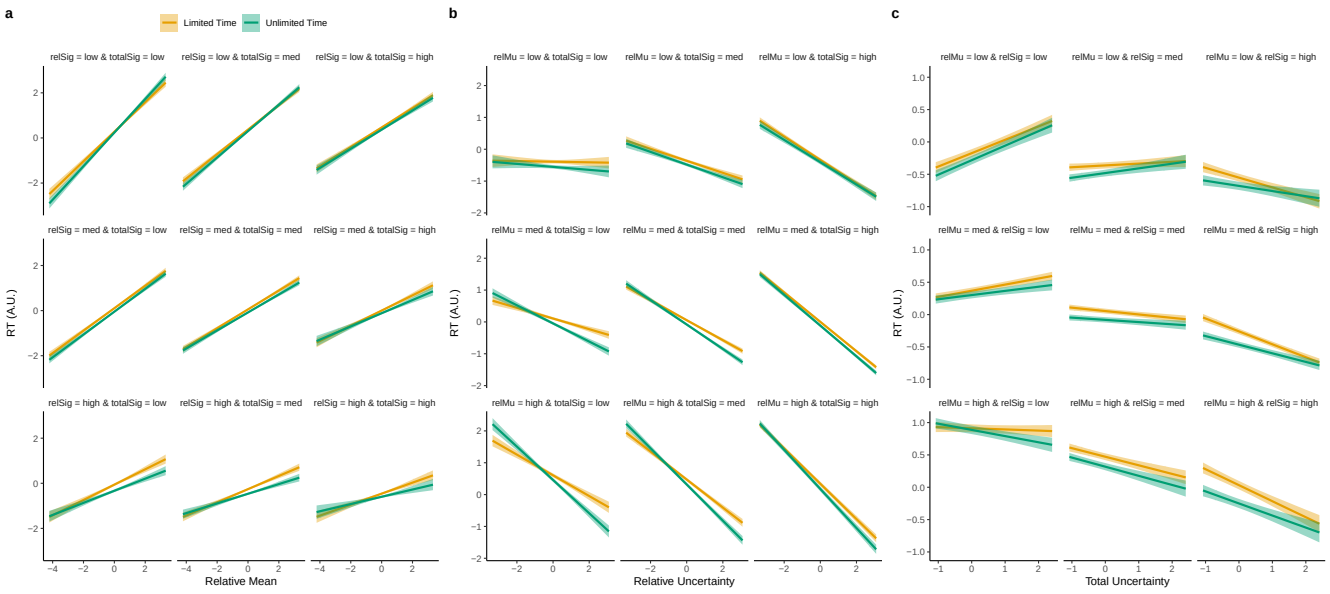

Figure S10. Four-way interactions for RT mixed effects regression. Four-way interactions of the RT Bayesian mixed effects regression illustrated in Fig. S8. Interactions are grouped in terms of relative means (**a**) and relative uncertainty (**b**). Continuous variables are split into discrete $[low, med, high]$ levels, based on $[mean - sd, mean, mean + sd]$, with relative uncertainty (relSig) increasing top to bottom (rows) and total uncertainty (totalSig) increasing from left ro right (columns). RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.

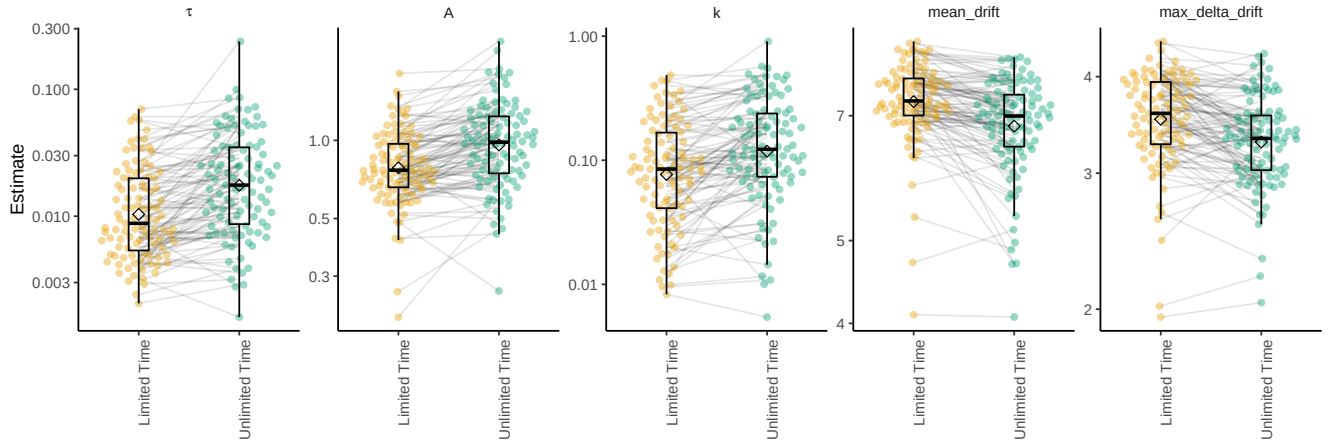

Figure S11. LBA parameters. Mean posterior parameter estimates of the LBA model, where each pair of connected dots is a single participant. Tukey boxplots show the group statistics, with the diamond indicating group means. τ is the non-decision time, A is the maximum starting evidence, k is the relative threshold, mean_drift is the average drift rate across all four options $\frac{1}{4}\sum_j v_j$, and max_drift_diff is the largest pairwise difference in drift rates $\max_{i \neq j} |v_i - v_j|$.

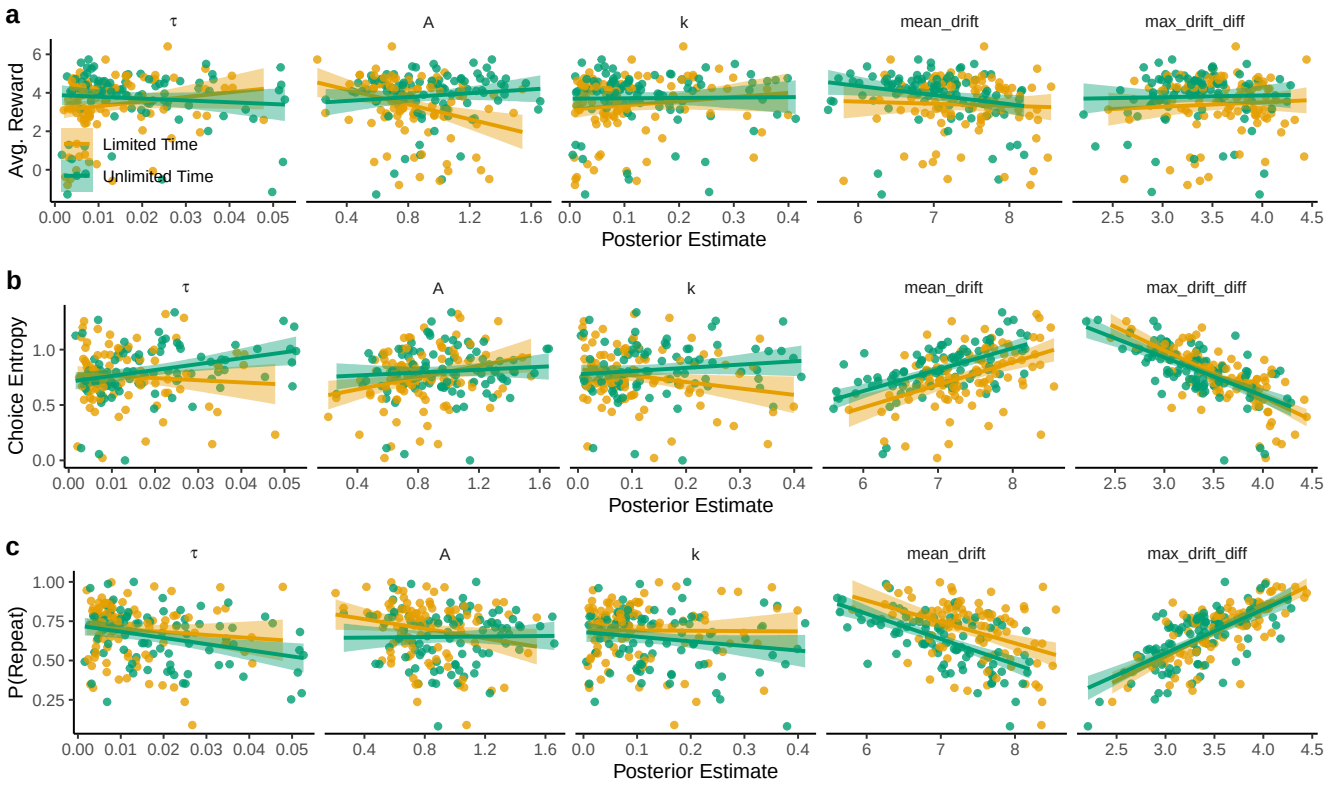

Figure S12. Comparison of LBA parameters and behavior. Each dot shows the mean posterior for each participant in each time condition, while the lines and ribbons are a linear regression and 95% CI. τ is the non-decision time, A is the maximum starting evidence, k is the relative threshold, mean_drift is the average drift rate across all four options $\frac{1}{4}\sum_j v_j$, and max_drift_diff is the largest pairwise difference in drift rates $\max_{i \neq j} |v_i - v_j|$. **a**) The only meaningful correlation between rewards and LBA parameters was found for maximum starting evidence A under time pressure ($r_\tau = -.25$, $p < .001$, $BF = 89$), where participants who were closer to making a decision prior to the start of a trial, earned lower payoffs. **b**) We find the strongest relationships between both drift rate variables and choice entropy, which were similar across time conditions. Participants with higher mean drift had more entropic choices (unlimited: $r_\tau = .37$, $p < .001$, $BF > 100$; limited: $r_\tau = .32$, $p < .001$, $BF > 100$), whereas participants with larger differences in drift rate were less entropic (unlimited: $r_\tau = -.49$, $p < .001$, $BF > 100$; limited: $r_\tau = -.57$, $p < .001$, $BF > 100$). We also find a weak correlation where higher maximum starting evidence was correlated with higher entropy for limited time rounds ($r_\tau = .15$, $p = .024$, $BF = 1.6$), and a moderate correlation where longer non-decision time corresponded to more entropic choices in unlimited time rounds ($r_\tau = .22$, $p = .002$, $BF = 18$). **c**) Similar to choice entropy, we again find the strongest relationship between the drift rate variables and the frequency of repeat choices, where higher mean drift produced less repeats (unlimited: $r_\tau = -.41$, $p < .001$, $BF > 100$; limited: $r_\tau = -.28$, $p < .001$, $BF > 100$), and larger differences in drift rate produced more repeat choices (unlimited: $r_\tau = .49$, $p < .001$, $BF > 100$; limited: $r_\tau = .58$, $p < .001$, $BF > 100$). We also find that higher starting evidence was correlated with more repeat choices in limited time rounds ($r_\tau = .58$, $p < .001$, $BF > 100$). In all plots, Tukey's fence has been applied to omit outliers for clearer visualizations, but all data are included in the statistical tests.

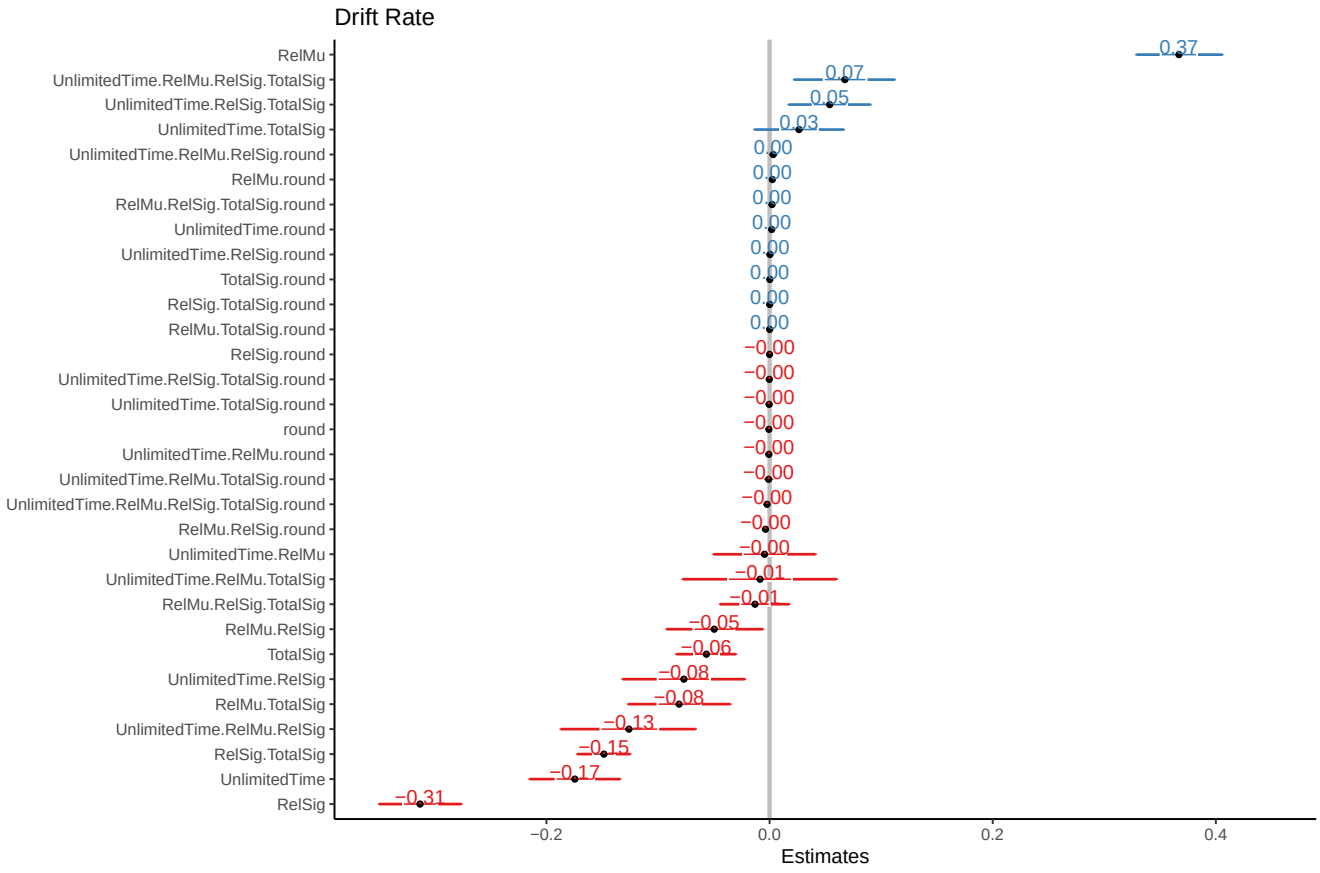

Figure S13. Coefficient plot of Drift Rate mixed effects regression. Posterior estimates of the Bayesian mixed effects regression predicting LBA drift rates. The mean posterior estimate is displayed numerically and indicated by the black dot, while the 95% HPD is illustrated by the length of the horizontal line. Coefficients are sorted by largest to smallest, with blue and red colors corresponding to estimates that are above or below 0, respectively, but do not indicate whether the difference is meaningful. See Figs. S14-S15 for interaction plots. RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.

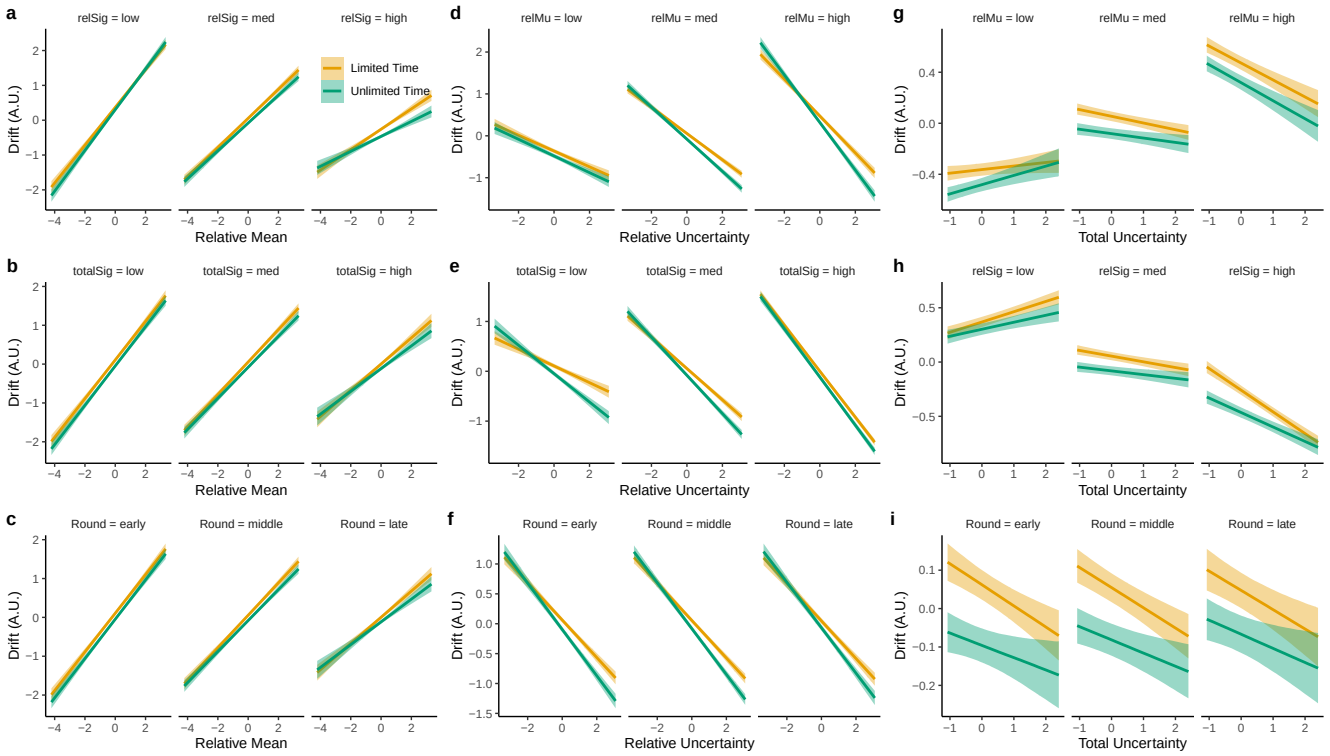

Figure S14. Marginal interaction plots for LBA mixed effects regression model. Marginal interactions of the LBA Bayesian mixed effects regression illustrated in Fig. S8. Interactions are grouped in terms of relative means (**a-c**), relative uncertainty (**d-f**), and total uncertainty (**g-i**). Continuous variables are split into discrete $[low, med, high]$ levels, based on $[mean - sd, mean, mean + sd]$. RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.

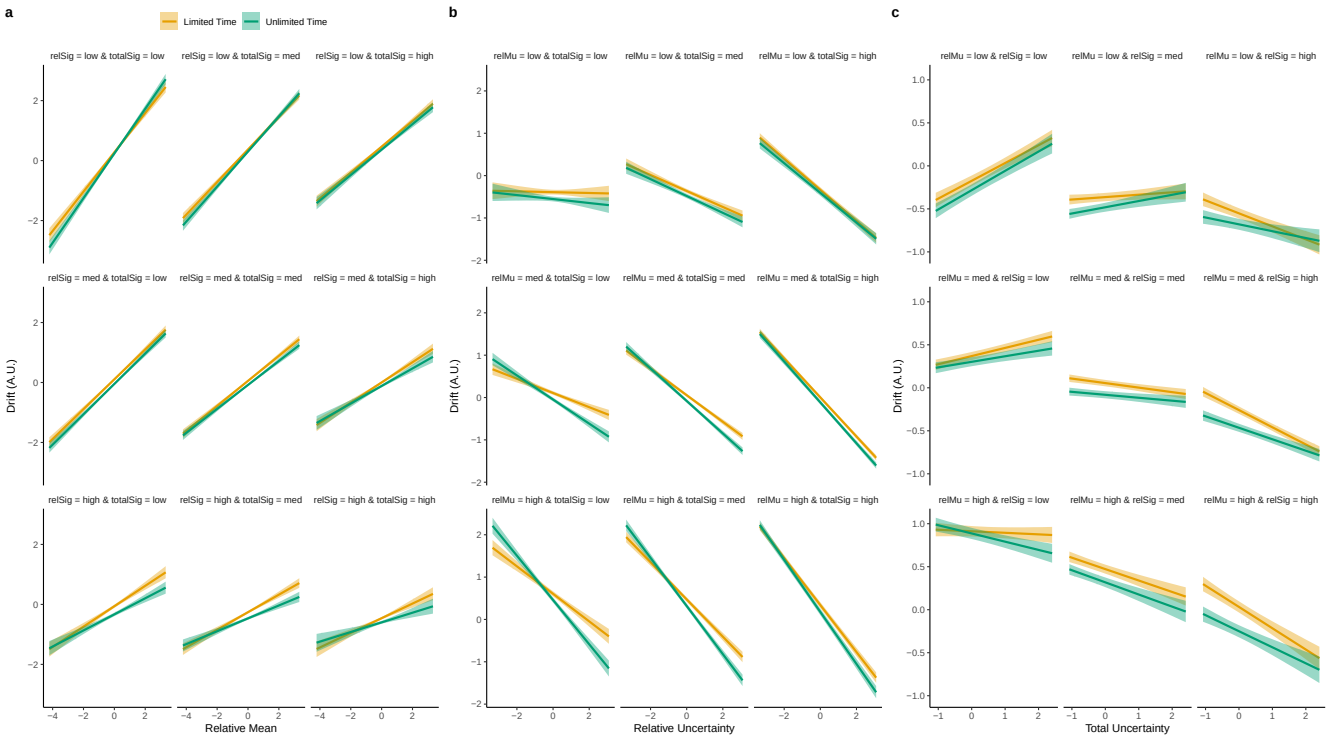

Figure S15. Four-way interactions for LBA mixed effects regression. Four-way interactions of the LBA Bayesian mixed effects regression illustrated in Fig. S13. Interactions are grouped in terms of relative means (**a**), relative uncertainty (**b**), and total uncertainty (**c**). Continuous variables are split into discrete $[low, med, high]$ levels, based on $[mean - sd, mean, mean + sd]$. RelMu: relative reward; RelSig: relative uncertainty; TotalSig: total uncertainty.