Supplemental Materials

Participant Inclusion Criteria

Because dual-task manipulations are interpretable only to the extent that participants are actually performing both, potentially interfering, tasks, we excluded the data of 11 participants whose concurrent task accuracy was less than 75% and the data of one participant who failed to meet a response deadline greater than 20 times. We employed a further step to remove participants who failed to demonstrate sensitivity to rewards in the decision task using second-stage choices, so as to detect reward task performance without biasing our primary comparisons of interest, which involve the first-stage choices. Namely, we excluded the data of 3 participants who repeated previously rewarded second-stage responses—i.e., P(stay|win)—at a rate less than 50%. The exclusion of the above participants does not affect the pattern of significance in the main strategy analysis described below: the WM-load lag×reward×transition terms (see Table 1) resulted in ps of .98, .48, and .008 for lag-0, lag-1, and lag-2 trials respectively.

Choice and RT Analyses

Mixed-effects logistic regressions were performed using the lme4 package (Bates & Maechler, 2009) in the R programming language. All coefficients were taken as random effects across subjects, and the estimates and statistics reported are at the population level. All predictors were coded as -1/1, with the exception of the lag-0, lag-1, and lag-2 indicator variables, which were coded as 0/1. Note that this model specification does not include a lag-2 main effect term because the combination of all three WM-load lag main effect terms is equivalent to the intercept term (which,

necessarily, needs to be specified in this class of linear model). Thus the lag-2 main effect term, of no particular interest to our hypotheses, is subsumed by the intercept term. To directly test differences between WM-load lag conditions (namely, the last three interaction terms in Table 1), planned comparisons were conducted using the esticon function (package doBy; Højsgaard & Halekoh, 2009) on the estimated model.

In the RT analysis, we included a number of binary nuisance variables in this linear model included motor responses on the last 5 trials (reflecting motor stay/switch costs), whether or not reward was obtained on the previous trial, and whether or not a correct Numerical Stroop response was made on the previous last trial (if applicable). These nuisance variables were entered in with a binary predictor indicating the WM-load lag, and contrasts between lag coefficients were calculated using the method described above.

Second-Stage Choice Analyses

According to our computational framework, model-based and model-free choice strategies make different choice predictions at the first stage of choice—on which our core analysis was focused—but not at the second stage. Accordingly, we expected no difference in second stage choice behavior across WM-load lags as the signatures of second-stage behavior should be indistinguishable under model-based versus model-free strategies. We specified a mixed-effects logistic regression model similar to the model used to analyze first-stage choices (see Results in main text) in order to analyze second-stage choices. However, unlike the main analysis, which examined first-stage stay/switch probabilities as a function of both the previous trial's reward and the previous transition type, this second-stage analysis did not include transition type as a factor because, unlike

for the first choice, the transition does not intervene between the second-stage choice and its reward, and thus has no causal relevance. Moreover, both RL models—by virtue of the Markov property—predict no effect of the transition type on second-stage choices. Accordingly, our analysis examined second-stage response repetitions (conditioned upon the response made on the subject's last visit to that second-stage state) as a function of WM-load lag and reward obtained on the subject's last visit to that state. The full model specification and resultant coefficient values are reported in Table S1. The critical pairwise contrasts were taken among the last three terms and are reported in the main text. Critically, we revealed no differences across the lag × reward interaction terms (all pairwise contrast *p*s > .28) confirming that second-stage behavior did not differ across WM-load lags.

Table S1. *Second-stage choice regression coefficients.*

Coefficient	Estimate (SE)	*p*-value
(Intercept)	0.57 (0.11)	< .0001
lag-0	0.16 (0.10)	0.089
lag-1	-0.16 (0.09)	0.067
lag-0 × reward	0.58 (0.11)	< .0001
lag-1 × reward	0.50 (0.08)	< .0001
lag-2 × reward	0.60 (0.08)	< .0001

Secondary Task Performance

The final group of included participants made Numerical Stroop judgments with an average accuracy of 86% in the WM task. We found no relationship between secondary task performance and task performance on WM-load trials (calculated as the total proportion of rewarded trials) arguing against the possibility of a more global

tradeoff between secondary task performance and choice task performance on WM-load trials (r=-.03, p=.90).

Reinforcement Learning Model

Our model follows closely the hybrid model described in Daw et al. (2011). The task consists of three states (first stage: s_A; second stage: s_B and s_C), each with two actions (a_A and a_B). The hybrid model consists of model-based and model-free subcomponents, both of which estimate a state-action value function $Q(s,a)$ that maps each state-action pair to its expected future value (cumulative reward). On trial t, we denote the first-stage state (always s_A) by $s_{1,t}$, the second-stage state by $s_{2,t}$, the first- and second-stage actions by $a_{1,t}$ and $a_{2,t}$, and the first- and second-stage rewards as $r_{1,t}$ (always zero) and $r_{2,t}$.

For the model free algorithm we used SARSA(λ) temporal difference learning (Rummery & Niranjan, 1994), which updates the value for the visited state-action pair at each stage i and trial t according to:

$$Q_{TD}(s_{i,t}, a_{i,t}) = Q_{TD}(s_{i,t}, a_{i,t}) + \alpha_i \delta_{i,t}$$

where

$$\delta_{i,t} = r_{i,t} + Q_{TD}(s_{i+1,t}, a_{i+1,t}) - Q_{TD}(s_{i,t}, a_{i,t})$$

is the reward prediction error (RPE), and α is a learning rate parameter. For the first-stage choice, $r_{1,t} = 0$ and the RPE is instead driven by the second-stage value, $Q_{TD}(s_{2,t}, a_{2,t})$; conversely at the second stage, we define $Q_{TD}(s_{3,t}, a_{3,t}) = 0$, since there is no further

value in the trial apart from the immediate reward $r_{2,t}$. The model uses an eligibility trace to propagate second-stage reward information to the first-stage values. Specifically, at the end of each trial, the first-stage values are updated according to:

$$Q_{TD}(s_{1,t}, a_{1,t}) = Q_{TD}(s_{1,t}, a_{1,t}) + \alpha_1 \lambda \delta_{2,t}$$

where λ is an eligibility trace decay parameter (Sutton and Barto, 1998). We assume that eligibility traces are reset to 0 between episodes (i.e., that eligibility does not carry over from trial to trial).

In general, a model-based RL algorithm works by learning a transition function (mapping state-action pairs to a probability distribution over the subsequent state), and immediate reward values for each state, then computing cumulative state-action values by iterative expectation over these. Specialized to the structure of the current task, this amounts to, first, simply deciding which first-stage action maps to which second-stage state (since subjects were instructed that this was the structure of the transition contingencies), and second, learning immediate reward values for each of the second-stage actions (the immediate rewards at the first stage being always zero).

We modeled transition learning by assuming participants used a Bayesian estimation scheme, starting with a uniform Beta prior over transition probabilities and updating using standard calculations for the Beta-Bernoulli family. Under this model, the estimated transition probability at time *t* is given by:

$$P(s_B | s_A, a_A) = (1 + N_{AB})/(2 + N_{AC} + N_{AB})$$

where N_{AB} denotes the number of times the participant observed a transition from state A to state B after taking action a_A.

At the second-stage (the only one where immediate rewards were offered), the problem of learning immediate rewards is equivalent to that for TD above, since $Q_{TD}(s_{2,t}, a_{2,t})$ is just an estimate of the immediate reward $r_{2,t}$; with no further stages to anticipate, the SARSA learning rule reduces to a delta-rule for predicting the immediate reward. Thus the two approaches coincide at the second stage, and we define $Q_{MB} = Q_{TD}$ at those states.

The model-based values are defined in terms of Bellman's equation (Sutton & Barto,1998):

$$Q_{MB}(s_A, a_j) = P(s_B|s_A, a_j) \max_{a \in \{a_A, a_B\}} Q_{TD}(s_B, a) + P(s_C|s_A, a_j) \max_{a \in \{a_A, a_B\}} Q_{TD}(s_C, a)$$

where we have assumed these are recomputed at each trial from the current estimates of the transition probabilities and rewards.

Finally, to connect the values to choices, we define net action values at the first stage as the weighted sum of model-based and model-free values $Q_{net}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1 - w)Q_{TD}(s_A, a_j)$ where w is a weighting parameter. At the second stage, $Q_{net} = Q_{MB} = Q_{TD}$. To accommodate our working memory load paradigm, we defined two different weights that operated on different trial types. We

define $w_{0/1}$ as the "lag-0/1" weight, which was used when working memory load occurred on the current or previous trial. The "lag-2+" weight w_2 was used on all other trial types.

We modeled choice probabilities as a softmax function of Q_{net}:

$$P(a_{i,t} = a|s_{i,t}) = \frac{\exp(\beta[Q_{net}(s_{i,t}, a) + p \cdot \mathrm{rep}(a)])}{\sum_{a'} \exp(\beta[Q_{net}(s_{i,t}, a') + p \cdot \mathrm{rep}(a')])}$$

where the inverse temperature parameter β governs the stochasticity of choices. The indicator function $\mathrm{rep}(a)$ is defined as 1 if a is a top-stage action and is the same one as was chosen on the previous trial, zero otherwise. Together with the "stickiness" parameter p, this captures first-order perseveration ($p > 0$) or switching ($p < 0$) in the first-stage choices (Lau and Glimcher, 2005).

In total, the algorithm contains 7 free parameters (β, α, $w_{0/1}$, w_2, λ, p), and nests pure model-based ($w = 1$, with arbitrary α_1 and λ) and model-free ($w = 0$) learning as special cases.

Experiment 2: Between-Subjects Conceptual Replication

Participants

A total of 89 undergraduates at the University of Texas were randomly assigned to one of two groups: the Single-Task (ST) condition and the Dual-Task (DT) condition. We used the same criteria for screening participants for adequate performance on both tasks as in Experiment 1. In particular, we excluded the data of 3 (2 DT) participants who failed to meet a response deadline greater than 15 times. To ensure that participants in the DT condition exhibited engagement with the secondary task, we excluded the data of 5

DT participants who exhibited a root-mean-squared-error on the tone counting task (detailed below) of 80 or greater. Following Experiment 1, we excluded the data of 6 (3 DT) participants who repeated previously rewarded second-stage response at a rate less than 50%. Consequently, 75 participants (37 DT) remained in the analyses.

Materials and Procedure

Both groups completed 200 trials of the two-step task using the same structure and stimuli as Experiment 1 in the main text. The DT condition followed the general tone-counting procedure of Foerde et al. (2006) but was modified to ensure that the concurrent task persisted over all stages of the decision task (Otto et al., 2011). This experiment used the same task flow and stimuli display as depicted in the No-WM-load condition of Figure 2. Both the ST and DT conditions followed the same trial timing procedure to ensure that, across conditions, a fixed amount of time elapsed each trial. In each stage, there was a 2-second response window and in the second stage, the outcome (a U.S. quarter or a zero) was presented for 1 second immediately at the conclusion of the response window.

In the DT condition, two types of tones, high-pitched (1000 Hz) and low-pitched (500 Hz) were played during each trial. Each trial stage was divided into 8 intervals of 250 ms, with tones occurring in intervals 3-10 (500-2,500ms after trial onset). The number of tones presented each trial varied uniformly between 1 and 4, occurring randomly within intervals 2-5. The base rate of high tones was randomly determined every 50 trials, varying uniformly between .3 and .7. Participants were instructed to maintain a running count of the number of high tones while ignoring the low-pitched

tones. At the end of each 50-trial block, participants reported their counts using the keyboard and were subsequently instructed to restart their count at zero.

Results and Discussion

We factorially examined stay probabilities in the same manner as Experiment 1, calculating first-stage stay probabilities as a function of previous reward and transition type between the ST and DT groups. Figure 6 reveals that ST participants exhibited a mixture of model-based and model-free strategies—mirroring the Lag-2 condition of Experiment 1 and Daw et al. (2011)—while DT participants appeared to rely only upon a model-free strategy. Critically, a mixed-effects logistic regression revealed a significant three-way interaction between WM load condition, previous reward, and previous transition type (full model specification and coefficient estimates are reported in Table 2). In other words, the cognitive demands imposed by concurrent tone-counting appeared to eliminate the influence of model-based strategy, reverting the DT participants to a putatively cognitively inexpensive model-free choice strategy. This complementary study, utilizing a fundamentally different design and concurrent task, corroborates the pattern of results seen in Experiment 1 and underscores model-based choice's reliance upon central executive resources.

References

Bates, D., and Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes Available at: http://CRAN.R-project.org/package=lme4.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Højsgaard, S., and Halekoh, U. (2009). doBy: Groupwise computations of summary statistics, general linear contrasts and other utilities Available at: http://CRAN.R-project.org/package=doBy.

Knox, W. B., Otto, A. R., Stone, P. H., and Love, B. C. (2012). The nature of belief-directed exploratory choice by human decision-makers. Frontiers in Psychology 2, 398.

Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav *84*, 555-579.

Rummery, G., and Niranjan, M. (1994). On-line Q-learning using connectionist systems.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (MIT Press).

Waldron, E. M., and Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. Psychonomic Bulletin & Review 8, 168-176.