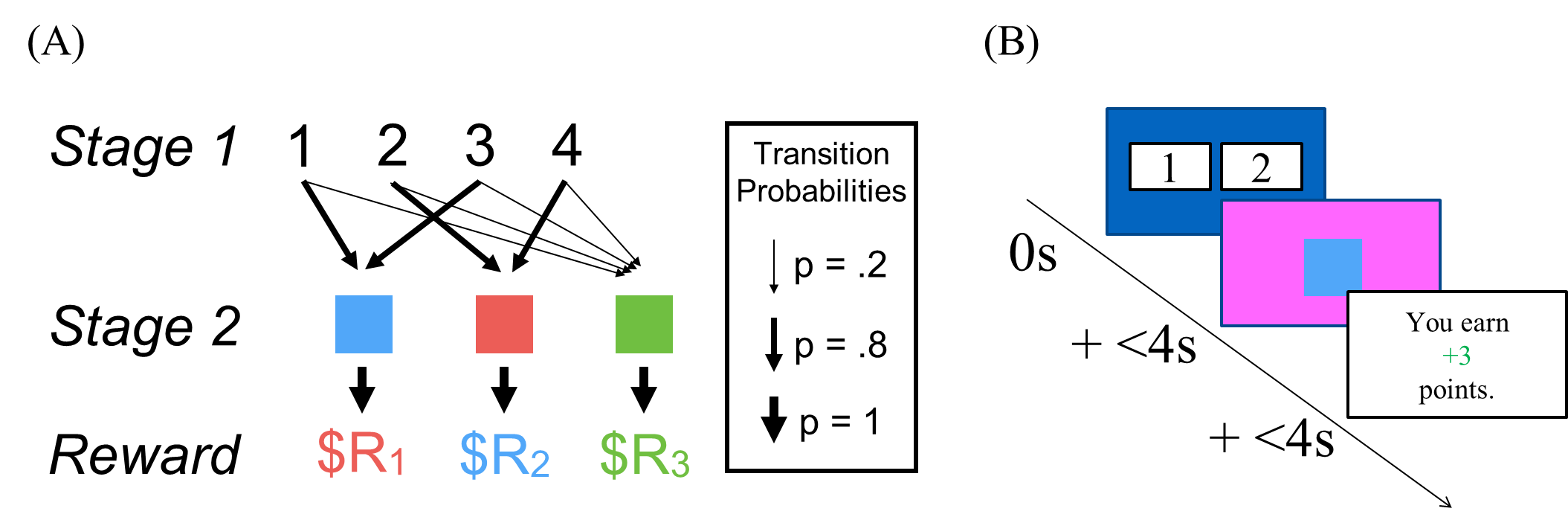
**Supplementary Materials**

**Experiment 1**

Methods

*Task design*

232 subjects were recruited on Amazon Mechanical Turk to participate in a two-stage Markov decision task. Four possible stage 1 options (represented by the numbers 1 through 4) each led probabilistically to one of three states (represented by the colors red, blue, and green). These states in turn had only one available action, which deterministically led to a reward. (See Figure 1.)

  
**Figure 1:** (A) Design of Experiment 1. Four number options lead probabilistically to one of three colors, which in turn lead to rewards. (B) Flow of the task. Subject is presented with two numbers, clicks on number 1, transitions to the blue square, clicks on it, and receives a reward.

On 80% of trials, numbers 1 and 3 led to blue and numbers 2 and 4 led to red. But, each had a 20% chance of leading to green. Subjects were explicitly told these transition probabilities, and were trained on them in the practice rounds. The high-probability transition of each number became the “goal” of that number – the goal of clicking on 1 would be to get blue, the goal of clicking on 2 would be to get red, etc.

On each trial subjects were presented with only two of the four number options. The option pairs to present were chosen randomly, with the constraint that the high-probability transitions of the two options had to lead to different colors – i.e. 1 and 3 could not be paired.

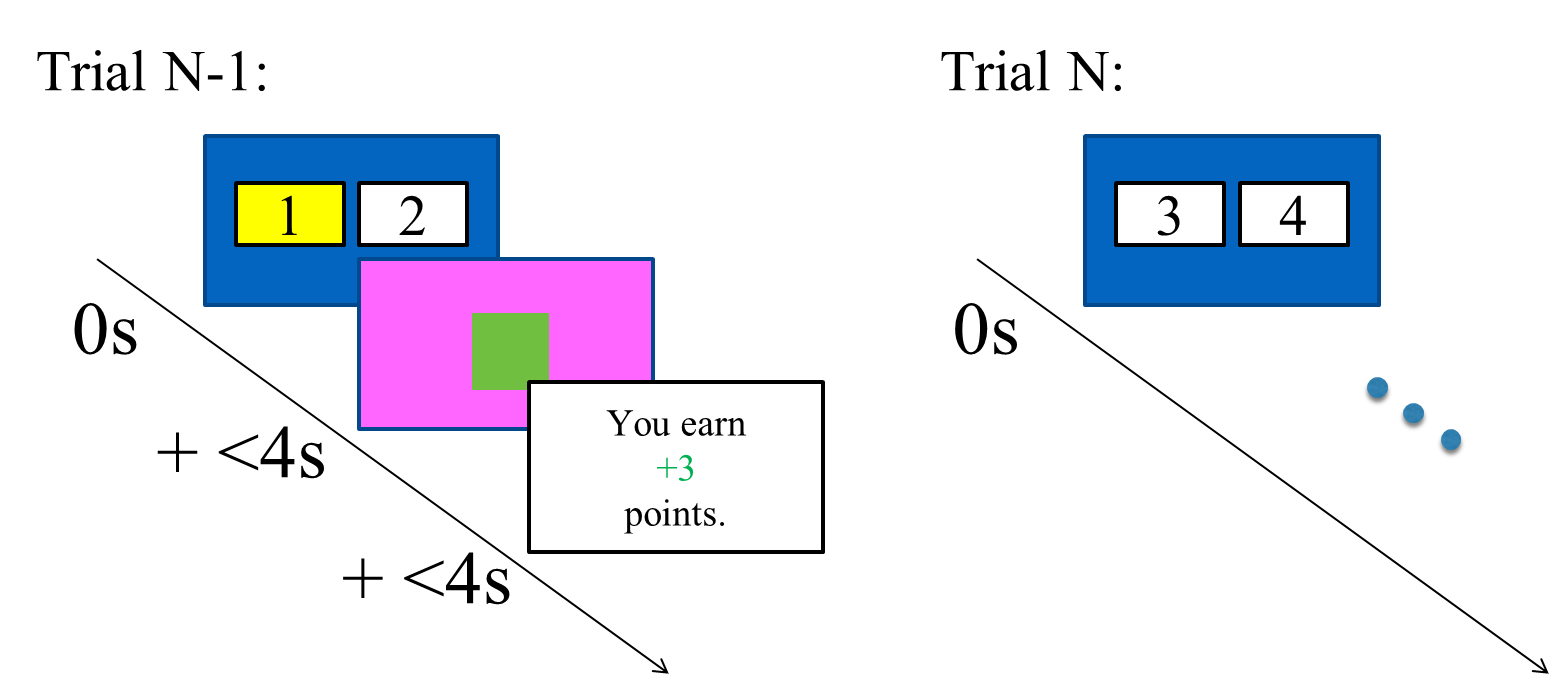
After clicking on one of the two numbers, subjects transitioned to a color, clicked on the color, and received a reward. The transitions from number to color were determined randomly according to the transition probabilities above. The rewards for each color were initialized uniformly at random on a range of -4 points to +5 points, and varied according to a bounded Gaussian random walk for the remainder of the experiment. After each round, the drift was sampled from a normal distribution with (μ=0, σ=1.8), rounded to the nearest integer, and added to the current reward level[[1]](#footnote-1).

Subjects completed 75 practice trials followed by 175 rewarded trials. On the rewarded trials, subjects had only 4 seconds to make their choice between the two numbers. If they did not make a choice within 4s the trial would time out and the next trial would begin. Subjects were excluded from analysis if they timed out on more than 50 trials.

Following Gläscher et al (2010), we also excluded subjects who did not meet a minimum threshold of learning. We ran a Monte Carlo simulation of 10,000 agents performing the task randomly, and determined the 95th percentile of their final scores. We excluded subjects whose final scores were below this cutoff.

*Congruent goal trials*

A critical subset of trials that we call *congruent goal* had two defining characteristics: they immediately followed a low-probability transition to green, and they did not present participants with the Stage 1 choice that they had chosen on the previous trial. For example, if the subject was presented with options 1 and 2, chose option 1, and transitioned to green, the next trial would be a congruent goal trial if it presented option 3 (paired with either 2 or 4). (See Figure 2.)

  
**Figure 2:** Congruent goal trial on trial N. On trial N-1, subject chooses option 1, transitions to green, and earns +3 points. The model-free goal value is +3. On trial N, the subject is presented with options 3 and 4. (Critically, option 3 has the same goal as option 1.)

Analysis

We restricted our analyses to congruent goal trials. We defined the “model-free goal value” as the reward obtained on the previous trial; that is, the most recent reward that immediately followed selection of the relevant goal. Although formal approaches to model-free reinforcement learning (e.g. Q learning) typically estimate value according to a geometrically-weighted sum of all past rewards (Sutton & Barto, 1998), past experimental research indicates more robust statistical estimates of model-free value assignment under the simplifying assumption that the most recent reward experience dominates value representation (Daw, personal communication). This estimation technique has been used in past studies of stochastic sequential decision-making paradigms (Daw et al 2011; Otto et al 2013). In order to further validate this analytic approach, below we show that it successfully recovers evidence for model-free value assignment to goal selection from the data generated by a formal computational model of our hypothesized mechanism.

As an initial, course-grained analysis, for each participant we computed the proportion of congruent goal trials on which they maintained their choice of goal, comparing trials on which model-free goal value was less than 0 (i.e. goal selection was followed by punishment) to trials on which the mean choice when it was greater than 0 (i.e. goal selection was followed by reward). We then compared these proportions across participants using a repeated measures t-test.

Second, as a more granular test, we regressed choice on the model-free goal value using a logistic mixed-effects model, estimating both random intercepts and random slopes at the subject level[[2]](#footnote-2).

Third, to definitively rule out any influence from a pure model-based or model-free system, we estimated a second mixed-effects model with approximate model-based and model-free action values as additional regressors. The model-based value of an action with a certain color goal was defined as the last reward that the subject received from that color. The model-free value of an action was defined as the reward received the last time the subject selected that action. In each case, convergence of our statistical model depended upon discounting reward values; we implemented a discounting parameter of .85 per trial.

These two values, model-based and model-free action, were computed for both available Stage 1 options for each critical trial. Then, the model-based value of the action which the subject did not choose was subtracted from the model-based value of the action which the subject did choose (in accordance with the coding scheme of the dependent variable), and the resulting single relative value became the model-based regressor in the mixed-effects model. The same procedure was applied to the model-free values. Therefore, the second mixed-effect model had three regressors: model-based, model-free, and model-free goal[[3]](#footnote-3).

We used the Wald test to derive the significance of the model-free goal regressor in the mixed-effects models. We also estimated null models (the full model with the model-free goal regressor removed), and performed both likelihood ratio tests and parametric bootstrap analyses to assess whether the model-free goal regressor increased the model’s likelihood enough to justify inclusion.

All mixed effect analyses were conducted in R (R Core Team, 2014), making use of the lme4 linear mixed effects package ([Bates and Maechler, 2010](http://www.sciencedirect.com/science/article/pii/S0896627311001255#bib7)).

Results

After applying our exclusionary criteria, there were 135 subjects and 3806 congruent goal trials. When the model-free goal value was positive, subjects chose the congruent goal 89.5% (SEM = 1.1%) of the time. When it was negative, subjects chose the congruent goal 68.6% (SEM = 1.7%) of the time. This difference was significant (paired t-test, t(134) = -12.5, p < .0001).

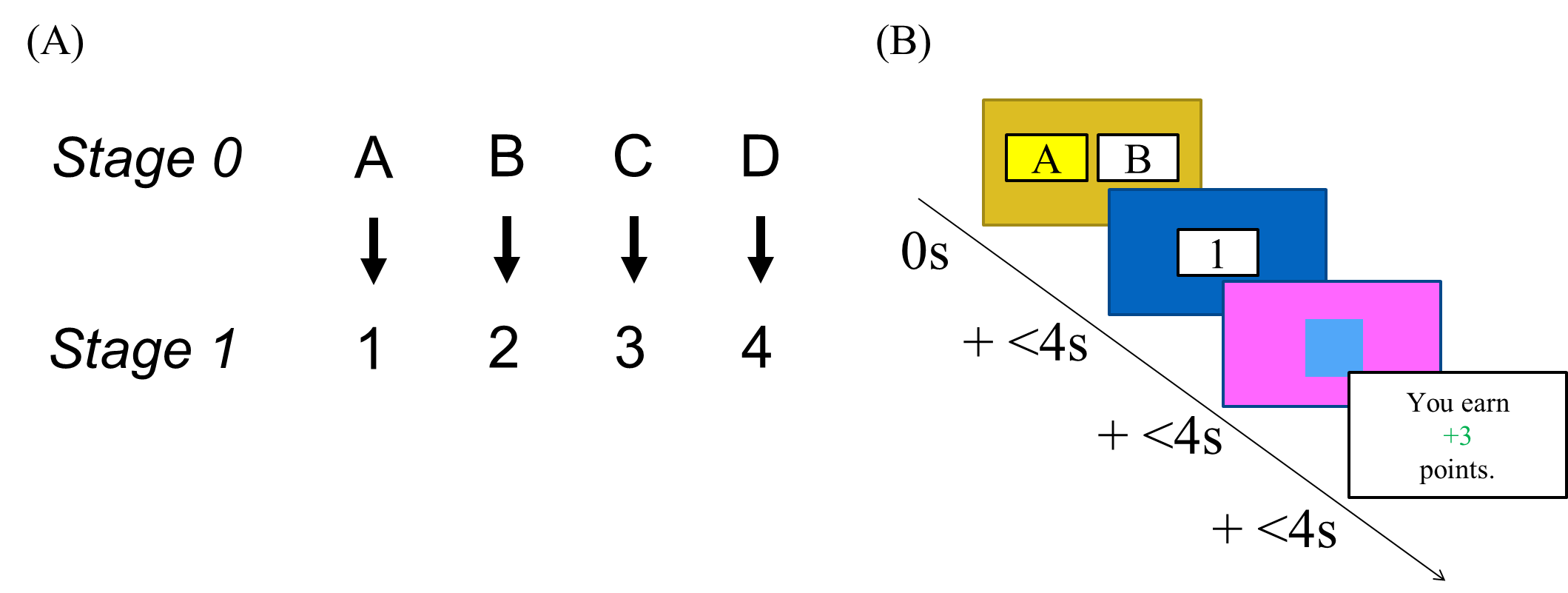
To estimate the models, we excluded 7 additional subjects who made the same choice on every congruent goal trial. In the simple mixed-effects model, the model-free goal regressor significantly predicted choice = .191, z = 12.1, p < .0001. The model was preferred to a null model without the reward (Likelihood ratio test, (2) = 266.0, p < .0001). In a parametric bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model.

In the complete mixed-effects model, with the model-based, model-free, and model-free goal values as regressors, the model-free goal regressor again significantly predicted choice = .200, z = 12.3, p < .0001. The model was preferred to the null model ((4) = 298.2, p < .0001). In a bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model. The model based predictor ( =.221, z = 7.3, p < .0001) was significant, and the model-free predictor ( = .054, z = 1.87, p = .062) predictor had a trending effect.

**Experiment 2**

Methods

312 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one above, with one change. Before being exposed to the structure of the main task, subjects were trained on a set of intuitive, deterministic transitions from letters to numbers (Figure 3A).

**Figure 3: (A) Transitions from Stage 0 letters to Stage 1 numbers. (B) Flow of congruent goal trials in Experiment 3. On these trials only, subjects choose between two letters and transition to a number, a color, and then a reward.

After becoming familiar with those transitions, subjects proceeded with the same task as above. All non-critical trials had exactly the same structure as in Experiment 1, with a choice between two numbers leading to a color, which in turn led to one of three drifting reward distributions. However, on congruent goal trials, subjects instead were presented with a choice between two letters. Subjects chose a letter and received a number (in accordance with the deterministic transitions in Figure 3A). They then clicked on that number and, in the usual way, got to a color which led them to a reward. The congruent goal trials thus required a goal-directed system to plan one extra step ahead (Figure 3B).

Analysis

All analyses were identical to those in Experiment 1.

Results

After applying our exclusion criteria, there were 173 subjects and 4755 congruent goal trials. After a reward, subjects chose the congruent goal 84.9% (SEM = 1.3%) of the time. After a punishment, subjects chose the congruent goal 68.8% (SEM = 1.6%) of the time. This difference was significant t(172) = -9.17, p < .0001.

To estimate the models, we excluded 8 additional subjects who made the same choice on every congruent goal trial. In the simple mixed-effects model the model-free goal regressor significantly predicted choice = .143, z = 9.62, p < .0001, and the model was preferred to a null model ((2) = 238.1, p < .0001; by bootstrapping, p < .001).

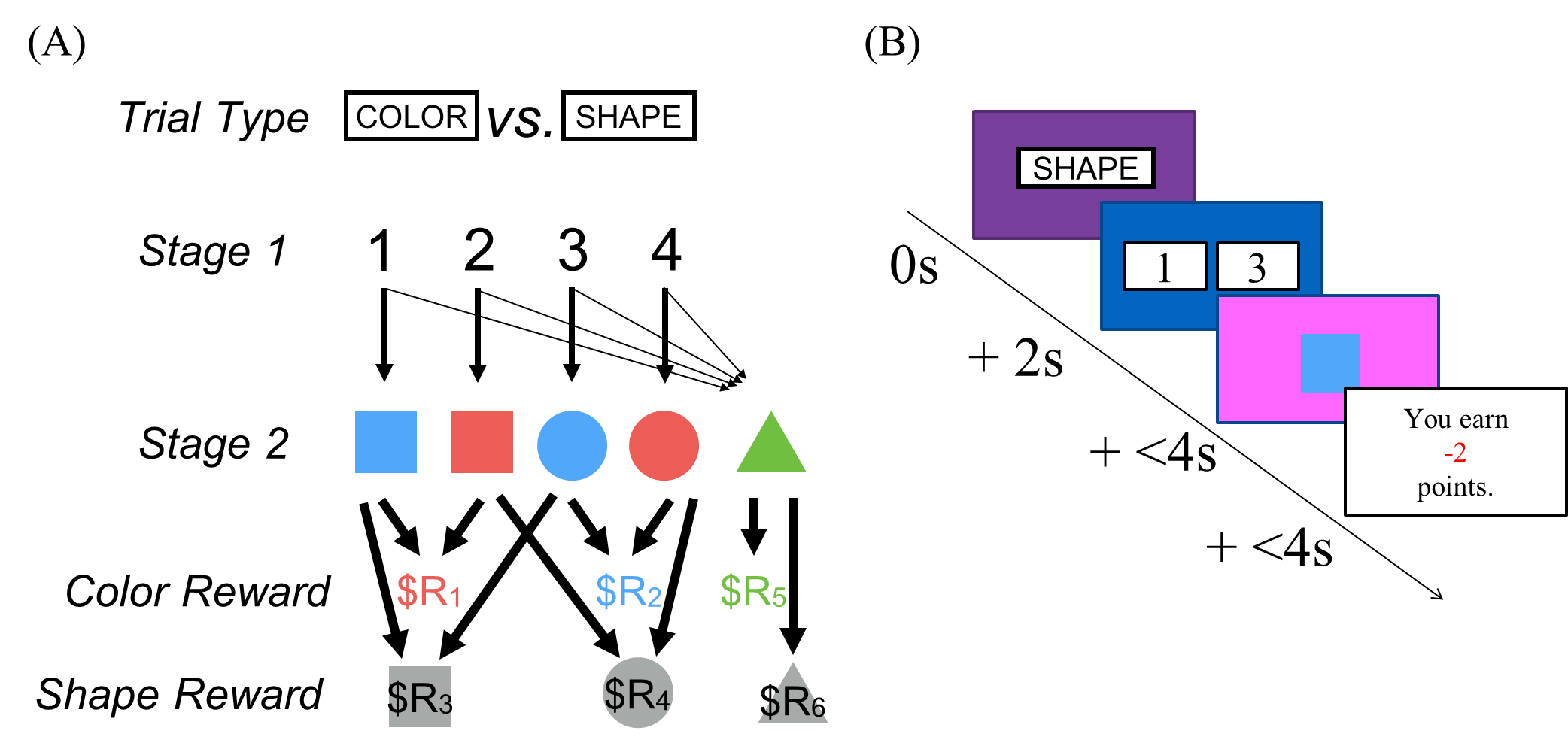
In the complete mixed-effects model the model-free goal regressor was significant, =.146, z = 9.56, p < .0001, and the model was preferred to a null model ((4) = 249.1, p < .0001; by bootstrapping, p < .001). Again, the model-based regressor ( = .149, z = 5.44, p < .0001) was significant, and the model-free regressor ( = .045, z = 1.65, p = .10) was trending.

**Experiment 3**

Methods

*Task design*

416 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one in Experiment 1, with the following change. In Experiment 1, Stage 2 states only varied in their color (blue, red, or green). In Experiment 3, they also varied in their shape. There were three shapes: square, circle, and triangle. See Figure 4.

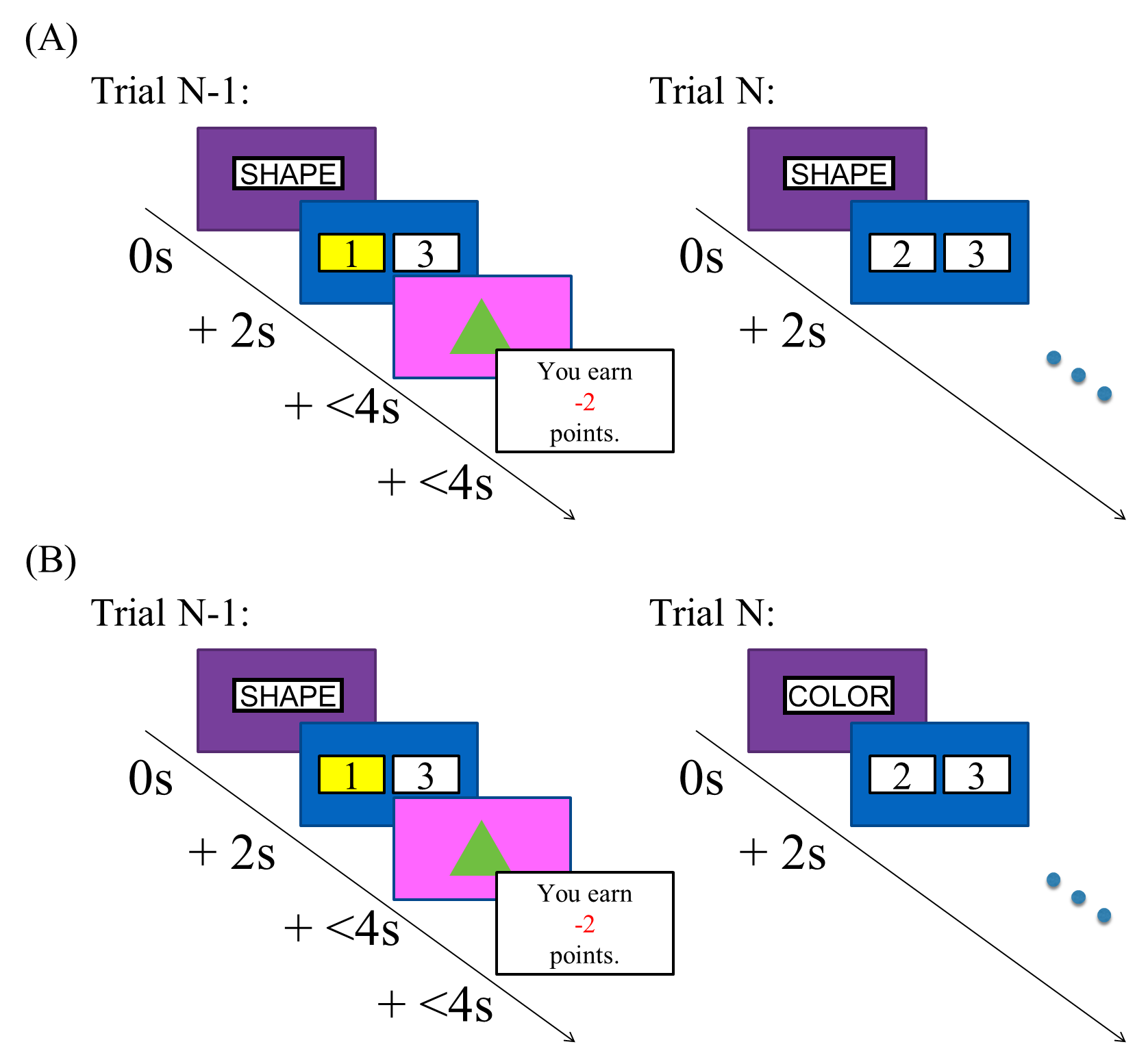
**Figure 4:** (A) Design of Experiment 3. Objects now differ in terms of color and shape. On color trials, the object color determines the reward; on shape trials, the shape does. (B) Task flow of Experiment 3. Subjects are told the trial type before proceeding with the trial.

In Experiment 1, reward distribution following Stage 2 was uniquely determined by the color of the Stage 2 option. In Experiment 3, each color and shape had a separate drifting reward distribution, and the reward value of an object could either be determined by its color or shape.

There were two trial types. On color trials, it was the color of the object which determined the reward. On shape trials, it was the shape of the object. Before each trial began, subjects were told the trial type. The flow of Experiment 3 is depicted in Figure 4B.

*Congruent and incongruent goal trials*

In Experiment 3, there were also two types of critical trials. In congruent goal trials, the trial type was the same as in the previous round[[4]](#footnote-4). In incongruent goal trials, the two trial types were different. (See Figure 5.)

  
Figure 5: (A) A congruent goal trial in Experiment 3. The trial types in the same in trials N-1 and N. (B) An incongruent goal trial. The trial types in N-1 and N are different.

Analysis

Our hypothesis predicts that the reward on the previous trial should predict subjects’ choices in congruent goal trials, not in incongruent goal trials. Therefore, we conducted the usual analyses on both congruent and incongruent trials separately, and then combined them into a new mixed-effects model with an interaction term between the model-free goal regressor and the type of critical trial.

Results

After applying our exclusion criteria, there were 303 subjects with 4231 congruent goal trials, and 2137 incongruent goal trials. On congruent goal trials, subjects chose the congruent goal 82.8% (SEM = 1.2%) of the time after a reward and 76.2% (SEM = 1.4%) of the time after a punishment. The difference was significant t(302) = -4.84, p < .0001. On incongruent goal trials, subjects chose what would have been the congruent goal 49.7% (SEM = 1.7%) of the time after a reward and 47.2% (SEM = 1.9%) of the time after a punishment. The difference was not significant t(282) = -.94, p = .35.

In the simple mixed-effect model on the congruent goal trials, the model-free goal regressor was significant = .056, z = 4.51, p < .0001. The model was preferred to a null model ((2) = 27.8, p < .0001; by bootstrapping, p < .001).

In the complete mixed-effect model on congruent goal trials, the model-free goal coefficient was again significant = .053, z = 3.97, p < .0001. The model was preferred to a null model ((4) = 27.3, p < .0001; by bootstrapping, p < .001). Likewise, the model-based = 0.313, z = 8.87, p < .0001 and model-free = .051, z = 2.04, p < .05 regressors were significant.

In the simple mixed-effect model on the incongruent goal trials, the model-free goal regressor was not significant = .009, z = .784, p = .433. The model was not preferred to a null model ((2) = .615, p = .74; by bootstrapping, p = .55). We also estimated a model with both congruent and incongruent goal trials, which included the model-free goal value and an interaction between that value and the trial type. In that model, the interaction was significant = .049, z = 2.62, p < .01, and the model was preferred to a null model with the interaction term removed ((4) = 10.7, p < .05; by bootstrapping, p < .01). Congruent goal trials were coded as 1 and incongruent goal trials were coded as 0, so the positive interaction term indicates that the model-free goal effect was significantly stronger for congruent goal trials.

**Simulations**

To validate our analytic approach we built a computational model of learning and choice including traditional model-based and model-free control, along with model-free goal learning. We used this computational model to generate simulated data for Experiment 3, and showed that our observed results were obtained if and only if the computational model included model-free goal learning.

Environment

The game was implemented as a Markov decision process with six states, the initial Stage 1 state and then one state for each Stage 2 object. State 1 had four possible actions (i.e. the four numbers), only two of which were available on any given trial. States 2-6 had only one possible action (i.e. clicking on the object), which led to a reward. The rewards were randomly generated for each agent by the same process as in the behavioral tasks.

Learning Mechanisms

*Model-free action value: SARSA*

The agents had three learning mechanisms. Their model-free reinforcement learning mechanism was the SARSA algorithm with eligibility traces (Sutton & Barto, 1998). Agents estimated a model-free value of the state-action pair (*s,a*), denoted *MFV(s,a)*. In Stage 1, agents chose an action *a* and transitioned to state *s.* The value update for *MFV(1,a)* occurs by temporal difference learning with learning rate :

In stage 2, agents chose the only available action *a’* (i.e. clicking on the object) and received reward *r*. Again, value update is given by temporal difference learning:

In addition, in keeping with prior computational models of stochastic sequential decision-making paradigms (Daw et al 2011; Otto et al 2013), we implemented an update of Stage 1 value representations following reward by applying an eligibility trace :

*Model-based action value*

Agents’ model-based learning mechanism implemented a dynamic programming technique. Agents maintained a model-based value of each state-action pair, denoted *MVB(s,a)*. We assumed knowledge of the trial-type-dependent reward distributions on the part of the model-based controller. Thus, we separately indexed Stage 2 states according to the relevant trial type. To calculate the model-based value of each action from state 1, agents estimated the transition probabilities of action *a* to state *s*, denoted by , by dividing the number of observed transitions from *a* to *s* by the total number of times *a* was selected. Then:

After transitioning to state *s*, performing the only available action *a’*, and receiving reward *r*, the model-based update was:

*Model-free goal selection*

The third learning mechanism was our proposed mechanism, model-free learning on goal selection. After a trial with chosen action *a* and received reward *r*, agents inferred the intended goal *g(a)* by:

Agents then updated the model-free value of the goal, *MFG(g(a))*, by:

*Action selection*

To determine the probability of an action *a*, agents took a weighted average *Wa* of the three values and entered it into a softmax function:

Thus, agents were characterized by five parameters: (the learning rate), (the eligibility trace), (the softmax temperature), (the model-based weight), and (the model-free weight).

Simulation Process

In each simulation, 200 agents were generated with randomly sampled parameters[[5]](#footnote-5). We ran two simulations: one where agents performed model-free goal learning, and one where they did not[[6]](#footnote-6). We then analyzed the agents’ behavior by the same process as in the behavioral tasks.

Results

In the simulation with model-free goal learning, on congruent goal trials agents chose the congruent goal 66.3% of the time after a reward and 51.2% of the time after a punishment (t(199) = -.694, p < .0001). The simple mixed-effects model on congruent goal trials estimated a model-free goal coefficient of .081 (z = 7.35,p < .0001), and was preferred to a null model ((2) = 63.1, p < .0001). The complete mixed-effect model showed similar results.

On incongruent goal trials, agents chose what would have been the congruent goal 48.6% of the time after a reward and 47.6% of the time after a punishment (t(186) = -.292, p = .77). The simple mixed-effect model estimated a model-free goal coefficient of .0098 (z = 0.649,p = .516), and was not preferred to a null model ((2) = .421, p = .81). The model that combined congruent and incongruent trials showed a significant interaction between model-free goal value and critical trial type ((3) = 14.4, p < .005).

In the simulation where agents did not perform model-free goal learning, agents showed no difference in behavior following a reward versus a punishment on congruent goal trials (t(199) = .71, p = .481). The simple mixed-effects model on congruent trials was not preferred to a null model ((2) = .483, p = .786). Neither was the complete model ((4) = 2.33, p = .675). The model combining congruent and incongruent trials did not show a significant interaction effect ((3) = 0, p = 1).

**Works Cited**

Bates, D., and Maechler, M. (2010). lme4: Linear mixed effects models using S4 classes. R package version 0.999375 33. <http://CRAN.R-project.org/package=lme4>.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, 0956797612463080.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

1. In cases where drift selected a reward level outside the bounds of [-4, 5], the reward would ‘rebound’ by the amount of the excess. For example, if the reward were at +3 and the drift were +5, the next reward would be . [↑](#footnote-ref-1)
2. There were several subjects in each experiment who made the same choice on every congruent goal trial. Because their random effects were impossible to calculate, these subjects were excluded from the model estimation. [↑](#footnote-ref-2)
3. All our regressors were grand mean centered. For convergence purposes, the models allow correlation among random slopes but not between random slopes and the random intercept. [↑](#footnote-ref-3)
4. Note that congruent goal trials here are equivalent to congruent goal trials in Experiments 1 and 2, because in those experiments all trials were color trials. [↑](#footnote-ref-4)
5. was sampled from a uniform distribution from 0 to 1, which we denote as U(0,1). was sampled from U(.5,1). was sampled from U(0,1.5). For the weights, three variables – ,,and – were sampled from U(0,1), and then and . [↑](#footnote-ref-5)
6. In the second simulation , forcing agents to use only model-based and model-free mechanisms. [↑](#footnote-ref-6)