Reward Associations Do Not Explain Transitive Inference Performance in Monkeys

Greg Jensen[1], **Yelda Alkan**[2,3], **Vincent P. Ferrera**[2,3,4], **and Herbert S. Terrace**[1,5]

[1]**Dept. of Psychology, Columbia University**
[2]**Dept. of Neuroscience, Columbia University**
[3]**Zuckerman Mind Brain Behavior Institute, Columbia University**
[4]**Dept. of Psychiatry, Columbia University**
[5]**Corresponding author (email: terrace@columbia.edu)**

ABSTRACT

The observation that monkeys make transitive inferences has been taken as evidence of their ability to form and manipulate mental representations. However, alternative explanations have been proposed arguing that transitive inference performance is based on expected or experienced reward value. We performed two experiments, in which we manipulated the amount of reward associated with each item in an ordered list, to test the contribution of reward value to monkeys' behavior in TI paradigms. Monkeys were presented with pairs of list items and were rewarded if they selected the item with the earlier list rank. When reward magnitude was biased to favor later list items, correct responding was reduced. However, monkeys eventually learned to make correct rule-based choices despite opposing incentives. The results demonstrate that monkeys' performance in TI paradigms is not driven solely by expected reward and that they are able to make appropriate inferences when given discordant rewards.

Keywords: Transitive inference, reinforcement learning, expected value, rhesus macaques.

INTRODUCTION

Knowledge that A>B and B>C can create knowledge that A>C, provided the ">" operator displays the transitive property. Knowledge that "A>C" is the result of transitive inference (TI), an effective problem-solving strategy, as for example, in sorting items, evaluating dominance hierarchies, and even the single-elimination logic of March Madness brackets.

Because of its seemingly logical character, TI was long assumed to require explicitly logical (and thus linguistic) faculties. It has therefore been a focus of research on the development of human cognition (see Wright 2001 for review). This conviction was powerful enough that, when the first demonstration of transitive inference in animals was published (McGonigle and Chalmers, 1977), the authors hedged with the title "Are monkeys logical?" Although transitive inference has since proved to be nearly ubiquitous among vertebrates (Jensen, 2017), the scope of the cognitive faculties that are necessary for some form of TI has correspondingly shrunk. In some species, TI is likely domain-specific (e.g. in fish it may be limited to assessment of dominance hierarchies) and thus potentially driven by associative mechanisms (Allen, 2006).

At one end of this spectrum are those who argue that TI in animals in general is merely the result of comparisons made among the associative strengths of various stimuli to rewards (reviewed in Vasconcelos, 2008). This view, is based on theoretical parsimony, insists that evidence of TI in animals can be explained as a model-free calculation of expected value that doesn't require cognitive representations or model-based calculations. The defense of this position is complicated by experimental designs in which only the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

adjacent pairs in an ordered list are trained (e.g. the pairs AB, BC, CD, and DE, drawn from the five-item list ABCDE). In this scenario, the stimuli B, C, and D each have an expected value of 0.5 during training (since they are correct for half of the trials in which they appear and incorrect for the other half). If, after training is complete, the novel pair BD (a "critical test pair") elicits performance above chance, then retrospective expected value is not sufficient. For example, the model-free algorithm "Q-learning," which makes choices using only retrospective expected value, cannot explain TI after training in which only adjacent pairs are presented. At the start of all-pairs testing, its responding is exactly at chance levels for critical test pairs (Jensen et al., 2015).

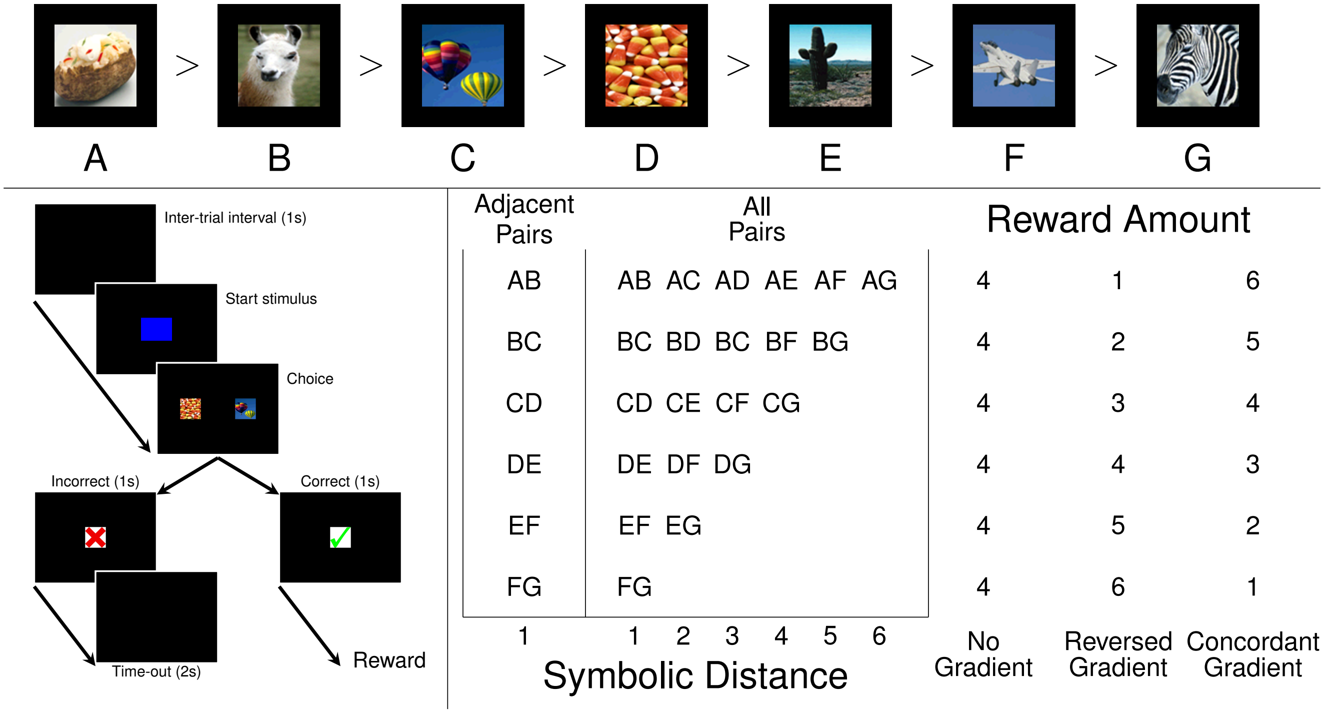
Failures like these motivated the development of more complicated associative models, such as Value Transfer Theory (VTT; von Fersen et al., 1991), which permits a propagation of expected value across stimuli from one pair to the next, without implementing a representation of the full list (Wynne, 1995). In effect, models like VTT perform a kind of prospective expected value calculation, despite being model free. Under some experimental designs, this yields above-chance performance on the critical test pairs.

Model-free and associative accounts of TI assume that any modification to the value of the training stimuli should impact performance. If, for example, a stimulus is a correct choice more often than it is an incorrect choice, then its value should grow. Thus, presenting DE more often than BC and CD should yield a corresponding boost to the value of D relative to B (since D is associated with rewards more often). Experimental tests of this hypothesis (Lazareva and Wasserman, 2012; Jensen et al., 2017) do not confirm this prediction. Subjects trained on adjacent pairs and then tested on the pair BD still favor B, even if subjects experienced D being paired with rewards far more frequently. Such results suggest that, although expected value calculations likely play some role, TI in non-human subjects reflects more than just comparisons of associative strength (Gazes et al., 2017).

Researchers still struggle to define the limits of "model-free" learning theories based on stimulus-response-outcome contingencies. Despite recent advances in machine learning (Mnih et al., 2015), model-free reinforcement learning is unable to account for human and monkey performance in implicit sequence learning paradigms (Jensen et al., 2015). A promising alternative relies on the concept of a cognitive map that represents spatial knowledge of the environment (Tolman, 1948; Epstein et al., 2017). Such representations can be combined with knowledge of item-item associations in "model-based" reinforcement learning.

At the cognitive level, however, the question remains, how is a representation used while executing a sequence? The advantage of a model-based representation of an ordered list is that rewards are treated as information about the representation's contents. Rather than merely pursuing maximal rewards without taking advantage of the context in which rewards are delivered, ordered representations allow the use of transitive reasoning, even when rewards are delivered unevenly. TI has provided some of the best evidence of cognitive representations in animals, specifically the representation of an ordered list, a type of one-dimensional cognitive map.

Despite demonstrations that reward expectation is not sufficient to explain TI in all cases, there is still ample evidence that the expected value of rewards can both influence behavior and yield measurable signals in the brain in serial tasks, whether list elements are presented pairwise (Pan et al., 2008) or simultaneously (Berdyyeva and Olson, 2011). The magnitude of these expected rewards may be inferred by transitive inference (Pan et al., 2014; Tanaka et al., 2015). Here, we tested monkeys' TI performance when individual list items were differentially rewarded, thereby creating gradients of reward size that were either concordant or discordant with how often a response to each stimulus was correct. This manipulation pits the TI rule ("choose the earlier list item") against a reward incentive ("choose the item that gives larger rewards"), thus revealing the contributions of experienced reward associations and cognitive rules to performance in a TI paradigm.



Adjacent Pairs	All Pairs	Reward Amount		
AB	AB AC AD AE AF AG	4	1	6
BC	BC BD BC BF BG	4	2	5
CD	CD CE CF CG	4	3	4
DE	DE DF DG	4	4	3
EF	EF EG	4	5	2
FG	FG	4	6	1
1	1 2 3 4 5 6 Symbolic Distance	No Gradient	Reversed Gradient	Concordant Gradient

Figure 1. Procedural details for Experiment 1 and 2. **(Top)** An example of a 7-item ordered list. **(Left)** Sequence of events for typical trial. Following an intertrial interval, each trial was initiated by touching a start stimulus. Upon a touch to either of the two presented stimuli, either a reward was delivered (along with a green check mark), or negative feedback (a red X). **(Right)** A diagram of the pairs presented during All Pairs sessions (Experiments 1 and 2) and Adjacent Pairs sessions (Experiment 2 only). The reward amount delivered varied according to the No Gradient condition (Experiment 1 only), the Reversed Gradient condition (Experiments 1 and 2), or the Concordant Gradient condition (Experiment 2 only).

EXPERIMENT 1: NO GRADIENT VS. REVERSED GRADIENT, ALL PAIRS

In a traditional study of TI, showing all possible pairs of list items yields a higher expected value for items earlier in the list, because they are paired with rewards in a larger proportion of trials. If indeed preference is driven chiefly by expected value, then success in this task ought to correlate with the degree to which the overall expected value for a stimulus is predictive of which response alternative in a particular pair is correct.

In order to see whether we could exploit this premise to undermine performance, we compared a standard procedure (where all rewards were of identical size) to one in which the size of the reward was larger for stimuli that were correct less often. Since items that were rarely correct yielded larger rewards, we called this manipulation a “reversed gradient,” that is, a gradient of delivered rewards that ran in the opposite direction than the gradient of the frequency with which a stimulus was a correct answer. Insofar as expected value drives behavior in pairwise choice settings, the introduction of the reversed gradient should undermine performance. Alternatively, performance that is guided by transitive comparisons of list items may be resilient to this reward manipulation.

Methods

Subjects

Subjects were four adult male rhesus macaques, N, O, R, and S. All subjects had prior experience with serial learning procedures, including transitive inference. However, subjects had not previously been exposed to manipulations of reward magnitude in the context of serial learning.

Subjects were housed individually in a colony room, along with approximately two dozen other macaques. Experiments were performed in their home cages, using the apparatus described below. To increase motivation, subjects were fluid-restricted to 300mL of water per day, or however much they were able to obtain by performing the task, whichever was greater. Typical performance yielded between 200mL and 300mL, whereas perfect performance could yield as much as 500mL. As needed, supplemental water was given to subjects each day after the end of the experimental session. Monkeys were also given a ration of biscuits (provided before experimentation each day) and fruit (provided after experimentation).

The study was carried out in accordance with the guidelines provided by the *Guide for the Care and Use of Laboratory Animals* of the National Institute of Health (NIH). This work, carried out at the Nonhuman Primate Facility of the New York State Psychiatric Institute, was overseen by NYSPI's Department of Comparative Medicine (DCM) and was approved by the Institutional Animal Care and Use Committees (IACUC) at Columbia University and NYSPI.

Apparatus

Subjects performed the task using a tablet computer. The tablet, running Windows 8.1, presented subjects with a 10.1" HD display (1266 x 768 resolution) which both presented stimuli and provided a capacitive touch screen interface to record responses. All tasks were programmed in JavaScript and run using the Google Chrome browser, set in kiosk mode.

In order to deliver rewards, the tablet was connected to a solenoid valve by way of an Arduino Nano interface, which opened the valve for fixed intervals when rewards were delivered via a steel spigot below the tablet. One "drop" of water corresponded to 0.25mL of fluid. When subjects received multiple drops, the valve opened and closed that many times in rapid succession to ensure that a consistent volume of liquid was being delivered. Unless otherwise noted, this device was identical to that described by Tanner et al. (2015).

This apparatus was mounted in a Lexan frame, which fit securely into the space created by opening the door to the subject's home cage. At the start of each trial, a solid blue square was presented in the center of the screen, in order to focus the subject's attention and to direct their hand toward a consistent center point. Touching it initiated the next trial. All experimental stimuli were 250x250 pixel images, presented to the right and left of the start stimulus.

Procedure

A fixed set of seven photographic images were arranged into an ordered list prior to the start of Experiment 1 (the list order is hereafter identified as ABCDEFG). This same set of stimuli was used for every session of Experiment 1, always retaining the same order. During each trial, two stimuli were presented simultaneously. The stimulus whose rank came earlier in the ordered list would always, if selected, yield a reward. The stimulus whose rank came later never yielded a reward. Subjects had to learn the ordering of the images by trial and error.

Each session consisted of up to 600 trials (fewer if the subject stopped responding before finishing the session). Each session was organized into "blocks" of 42 trials each. During a block, each of the 21 possible pairings of the seven stimuli were presented twice, once for each spatial arrangement.

Sessions consisted of one of two conditions. In the "No Gradient" condition, every correct response was rewarded with four drops. In the "Reverse Gradient" condition, a correct response earned a number of drops equal to its rank. In all cases, an incorrect response yielded no reward. So, for example, when presented with the pair AB, a response to A earned 1 drop (because A has a rank of 1), whereas a response to B would earn no drops (because it is incorrect). However, when presented with the pair FG, a response to F would earn 6 drops (because its rank is 6), and a response to G would earn no drops. As a result, the

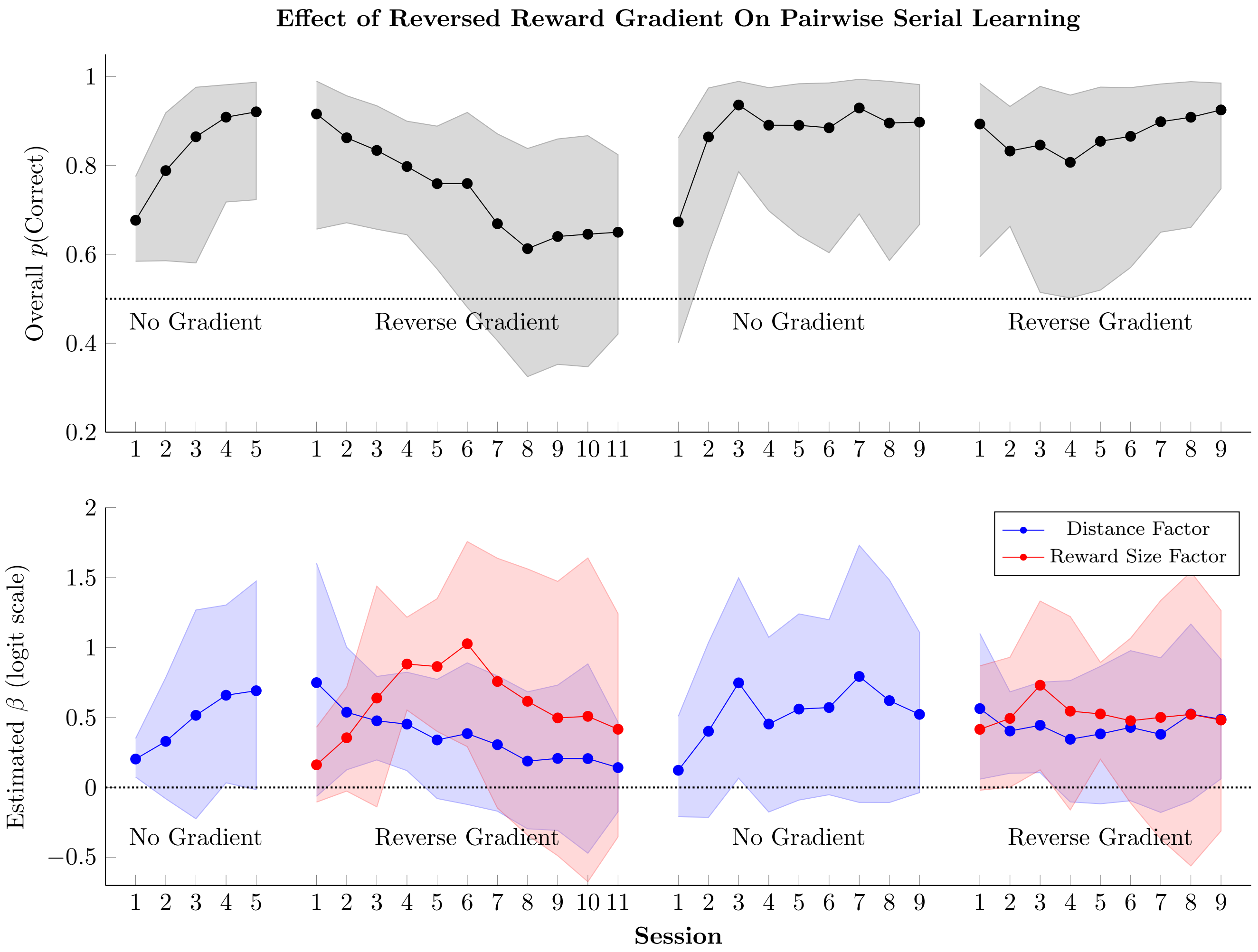


Figure 2. Population-level estimates of session-by-session performance in Experiment 1. Shaded regions correspond to the 95% credible interval. (Top) Estimated response accuracy for an average subjects, presented on a logit scale. (Bottom) Estimated population parameters for the distance parameter (β_D) in blue and the reward parameter (β_R) in red.

stimuli that were correct less often are worth more in the cases when they are correct (see Figure 1 for details).

Subjects completed five sessions of the No Gradient condition in order to learn the ordered list. They then completed 11 sessions of the Reverse Gradient condition using the same set of stimuli. Next, they completed 9 sessions in the No Gradient condition, and finally they completed 9 sessions in the Reverse Gradient condition.

Results

Performance was modeled on a session by session basis using binomial regression, with a logit link. In each model, an intercept term $\beta_\emptyset$ was included. In the No Gradient condition, performance was predicted as a function of the symbolic distance between pairs, yielding one additional slope term β_D. In the Reversed Gradient condition, performance was predicted in terms of both symbolic distance and the reward magnitude of the correct alternative, yielding two additional slope terms β_D and β_R. These parameters were fit using multi-level models, yielding both population- and subject-level estimates, which were implemented in the Stan programming language (Carpenter et al., 2017). The code for these models

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

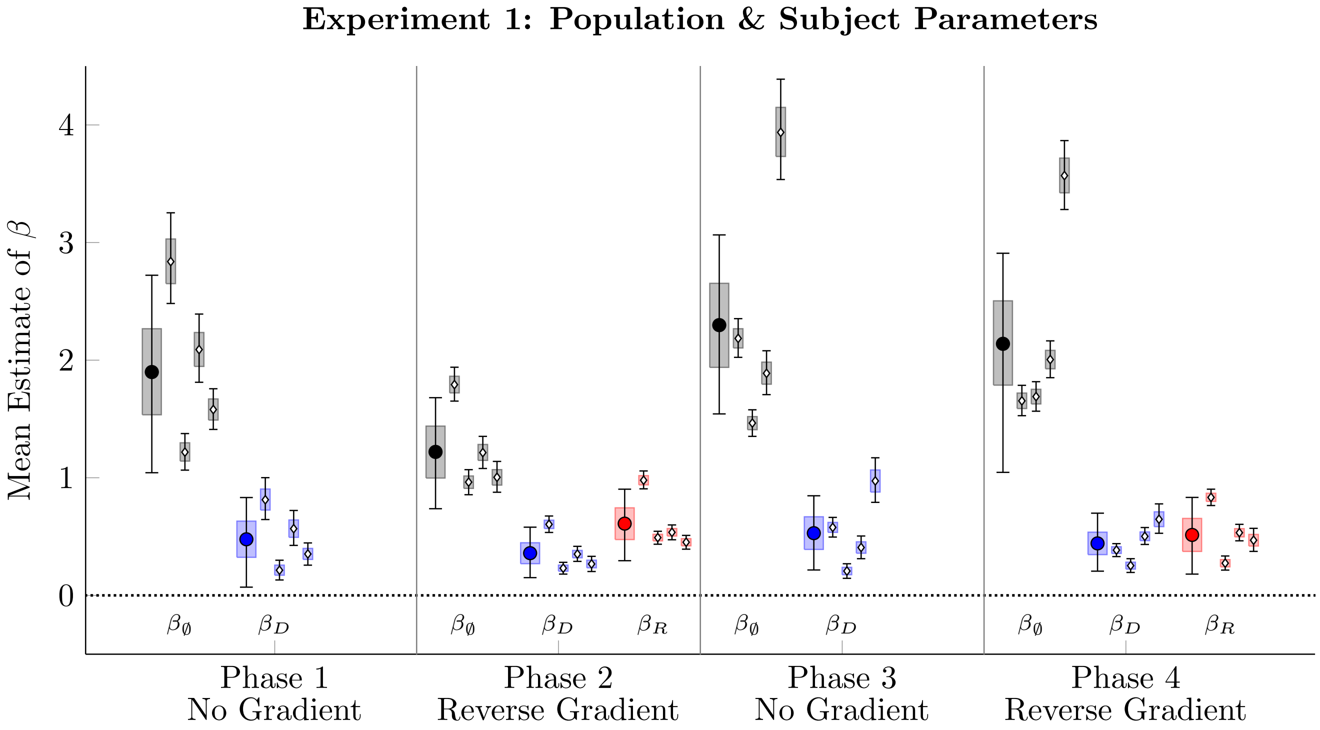


Figure 3. Mean population- and subject-level estimates of parameters in Experiment 1, presented on a logit scale. Filled circles correspond to population estimates, and white diamonds correspond to subject-level parameters for each subject. Whiskers correspond to the 99% credible interval, and boxes correspond to the 80% credible interval.

is provided in the electronic supplement.

Figure 2 (top) depicts the overall response accuracy of an average subject in each session. During the first five sessions, response accuracy steadily rose to around 90%. However, when the reverse gradient was introduced, performance remained high in the early sessions, but gradually degraded. When the reward gradient was removed in the third phase, performance returned to ceiling levels. Finally, in the fourth phase, the reversed gradient was reintroduced. This time it no longer resulted in poorer performance. By the end of the fourth phase, it was as high as it had been in the No Gradient conditions.

The functions shown in Figure 2 (bottom) place these results in clearer perspective. As response accuracy grew in the first phase, so did the size of the symbolic distance effect. However, as performance deteriorated in the second phase, the distance effect shrank. A large reward effect appeared in its place, but it gradually shrank during the second half of that phase. In the third phase, in which the reversed gradient was removed, the symbolic distance effect reappeared . Finally, in the fourth phase, both effects held steady at moderate levels.

Figure 3 plots the means of these regression parameters for each phase at both the population level (points) and at the level of individual subjects (diamonds). From these plots, it is clear that both symbolic distance and reward contributed equally in both Phases 2 and 4. In but Phase 2, however, the intercept was lower (thus leading to lower performance globally). By contrast, the intercept in Phase 4 was comparable to that seen in the preceding No Gradient phase.

The implications of the parameters in Figure 3 are more fully depicted in Figure 4, which plots the model's estimated response accuracy for each of the 21 stimulus pairs (boxplots), as well as the observed mean accuracy of subjects for each pair (points). These estimates were calculated from the posterior parameter distributions of our regression model. In both Phase 2 and Phase 4, the reward gradient produces

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

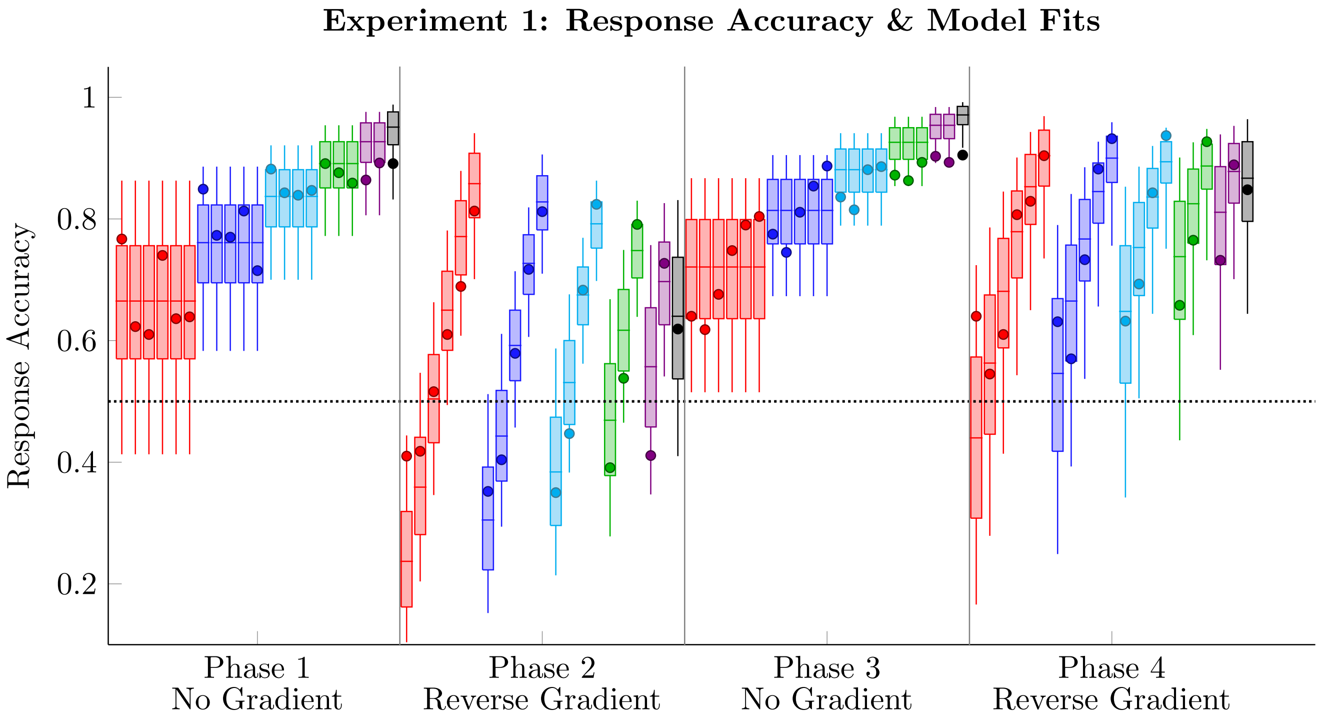


Figure 4. Response accuracy for each pair in each phase of Experiment 1. Points depict the means subject response accuracies, as computed directly from the data. Box-and-whisker plots correspond to model estimates of response accuracy, with whiskers corresponding to the 99% credible interval and boxes corresponding to the 80% credible interval. Pairs are sorted by distance ($D = 1$ in red, $D = 2$ in blue, $D = 3$ in cyan, $D = 4$ in green, $D = 5$ in purple, and $D = 6$ in black,), then by rank (AB, BC, CD, etc.).

a clear distortion in response accuracies, pushing the pair AB to chance levels or below (despite the fact that stimulus A was rewarded every time it was selected). However, despite this distortion, overall accuracy in Phase 4 was also higher overall, yielding reliable performance despite the distorting effect of the reward gradient.

Discussion

Experiment 1 demonstrates clearly that reward magnitude influences performance in a pairwise serial learning task. Despite extensive prior experience with the single list of stimuli used in this experiment, subjects went from average performance on the pair AB of nearly 80% in Phase 1 to merely 40% in Phase 2. Had subjects simply ignored the magnitude of rewards and focused on the frequency of reward delivery, then they could presumably have retained the 90% overall accuracy seen at the end of Phase 1.

It is somewhat less clear, however, that the reversed gradient disrupted performance in the final phase. The reduced overall response accuracy seen in Phase 2 was not seen in Phase 4. Although a distortion by the reward gradient was evident, a high baseline accuracy was maintained.

Associative models predict that varying reward magnitude should change behavior. Furthermore, it is also clear from Figure 4 that the Reversed Gradient condition imposed a distortion on performance. What is not immediately obvious is whether the sort of distortion observed in Experiment 1 is consistent with an associative account. To see why, consider the following table that describes the expected value of each alternative:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

Stimulus	Proportion of pairs in which choosing the stimulus is rewarded	Reward size	Expected value (proportion · reward)
A	6/6	1 drop	1 drop
B	5/6	2 drops	1.667 drops
C	4/6	3 drops	2 drops
D	3/6	4 drops	2 drops
E	2/6	5 drops	1.667 drops
F	1/6	6 drops	1 drop
G	0/6	0 drops	0 drops

Contrary to expectation, when all 21 pairs are presented during training, the nominal distribution of expected values is not monotonic, nor is it flat. Instead, it predicts that the stimuli C and D should be tied for the highest expected value, that B and E be tied for a somewhat lower value, and that A and F be tied at a value that is lower still. G has the lowest expected value, since it was never rewarded. If choice is then based on a comparison of these overall expected values, then this calculation implies that the pairs AB, AC, AD, AE, BC, and BD should all be below chance. Meanwhile, AF, BE, and CD should be selected at chance levels, and all remaining pairs should be above chance. Broadly, the associative model does not merely predict a distortion, but also predicts where the midpoint of that distortion should be: around the AF, BE, and CD pairs.

This pattern of predicted above- and below-chance responding closely corresponds to the performance observed in Phase 2. Despite extensive prior experience with serial learning procedures, and extensive training on this 7-item list, subjects in Phase 2 appeared to be fulfilling the associative prediction. However, the predicted pattern was not observed in Phase 4. Although the relative biases seen in Phase 4 were the same as those seen in Phase 2, performance on all pairs was at or above chance. Thus, although an associative mechanism could plausibly be influencing performance in Phase 4 such a mechanism does not entirely accounts that performance.

What is especially interesting about this result is that the time-course implied by Figure 2 suggests that subjects in Phase 2 experienced a kind of erosion to their prior training. By whatever mechanism, subjects had “learned the list” in Phase 1 to a high degree, and maintained high performance in the initial trials of Phase 2. Only gradually did performance begin to drop, doing so steadily until the onset of Phase 3. This may indicate that the culprit for poor performance in Phase 2 acts not during learning, but during the consolidation that unfolds over intervening days.

Whatever mechanism it was that gradually drove Phase 2 performance toward the reward association prediction, it only partially re-emerged in Phase 4. This could suggest either of two possibilities. On the one hand, it could be that the brain is simultaneously updating an associative representation of value and a cognitive representation of stimulus order, and that behavior is influenced by both systems. The improved performance in Phase 4 would therefore be a sign of subjects giving greater credence to their serial representations, and less credence to expected value. However, another possibility is that “expected value” is calculated by a function that can weigh “probability of being correct” separately from “reward amount.” In this case, performance in Phase 4 could still be accounted for by entirely associative mechanisms, so long as a means exists to parametrically attenuate the contrasts between reward amounts. Experiment 2 evaluates that possibility.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

EXPERIMENT 2: REVERSED VS. CONCORDANT GRADIENT, TRANSITIVE INFERENCE

As noted above, the results of Experiment 1 may still be explainable in terms of associative mechanisms, given some mild constraints on the form such models can take. Because it does not confirm the cognitive contribution to explaining performance, it also doesn't provide any insight into the question of whether transitive inference is undermined by reward gradients. Because all pairs were being presented at all times in Experiment 1, TI was at no point necessary to explain performance. That is, although TI may have been occurring, the experiment did not provide a critical test of this hypothesis. For example, performance throughout Experiment 1 could be attributed to the memorization of which response was correct for each pair.

With this in mind, we performed a second experiment, in which the reversed gradient was applied, but only adjacent stimulus pairs (AB, BC, CD, etc.) were presented during training. When all pairs were subsequently presented during a testing phase, critical test pairs such as BD were novel.

As a contrast against this condition, Experiment 2 also introduced a "Concordant Gradient" condition, which aligned rewards with expected frequency of reward, rather than reversing it (e.g. choosing A would yield 6 drops if correct, choosing B would yield 5 drops if correct, and so forth). Insofar as the reversed gradient appeared to interfere with all-pairs performance, a concordant gradient would be expected to facilitate such performance.

Methods

Subjects & Apparatus

These were identical to those in Experiment 1.

Procedure

In order to rely on transitive inference, without a reward confound, Experiment 2 was divided into training phases (during which only adjacent pairs AB, BC, CD, DE, and FG were presented) and test phases (during which all 21 pairs were presented). Between 5 and 9 sessions of training were then followed with 2 to 3 sessions of testing. Sessions lasted up to 1000 trials.

Sessions consisted of one of two conditions. The first was the "Reverse Gradient" condition, identical that described in Experiment 1. The second was the "Concordant Gradient" condition, in which stimuli of earlier rank yielded larger rewards. Such a gradient is "concordant" in that the reward associated with a stimulus is positively correlated with the odds of that stimulus being correct in a random pairing. Correct responses to A yielded 6-drop rewards, correct responses to B yielded 5-drop rewards, and so on, until correct responses to F yielded 1-drop rewards.

At the start of each training phase, a new list of seven unfamiliar stimuli was used. Between 5 and 9 sessions of training were then followed with 2 to 3 sessions of testing. This was collectively considered a "training cycle." Subjects learned a total of 12 lists over the course of the experiment.

Subjects first learned four lists under the Reverse Gradient condition (each having a training phase and a testing phase). This was followed by two lists learned under the Concordant Gradient condition. After these six lists were completed, subjects took a one-year break from the experiment (during which they participated in experiments that had no differential gradients of reward magnitude). Following this break, they learned four lists using the Concordant Gradient, followed by two lists using the Reverse Gradient. The purpose of the break was to provide a control against order of learning effects, since it was unclear whether a Reverse-to-Concordant transition would result in similar overall performance than a Concordant-to-Reverse transition.

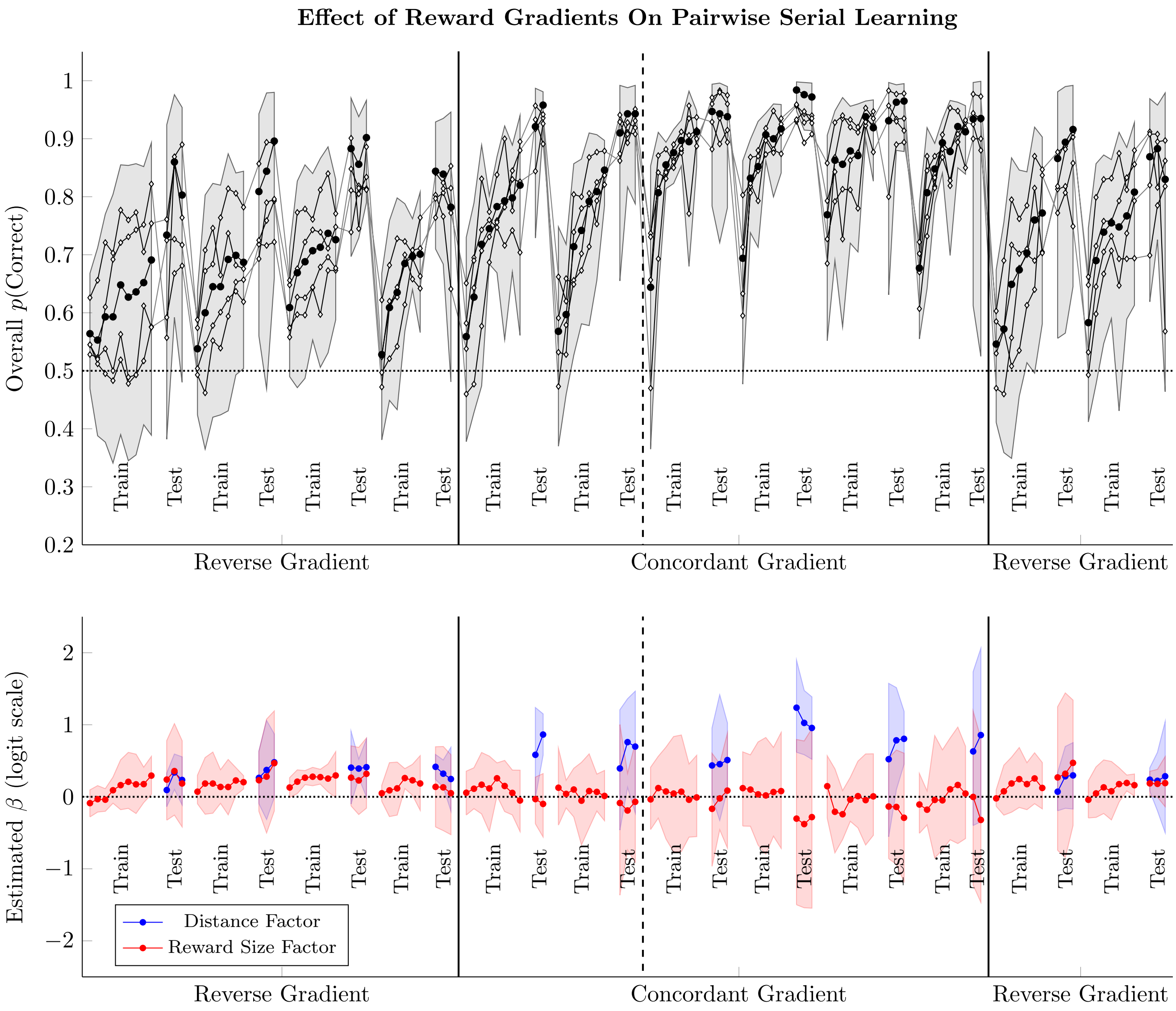


Figure 5. Population-level estimates of session-by-session performance in Experiment 2. Shaded regions correspond to the 95% credible interval. (Top) Estimated response accuracy for an average subjects, presented on a logit scale. (Bottom) Estimated population parameters for the distance parameter (β_D) in blue and the reward parameter (β_R) in red.

Results

As in Experiment 1, performance was evaluated on a session-by-session basis using binomial regression, implemented as multi-level models using the Stan programming language. During adjacent-pair training, the reward gradient was used as a predictor (yielding a β_R term), but because all pairs had the same symbolic distance of 1, no distance term was included. Both the β_D term and the β_R term were used to fit models of the all-pairs testing phases.

Figure 5 plots session-by-session overall response accuracy and empirical means from each subject (top panel) and the mean population slope parameters (bottom panel) during four training-testing cycles of the Reversed Gradient condition and two cycles of the Concordant condition. Each of these twelve cycles of training and testing made use of a novel ordered list of stimuli. In each cycle, response accuracy clearly jumped at each transition from training to testing, with positive values for the β_D parameter in each case, both classic signs of transitive inference. These effects appeared larger overall in the Concordant

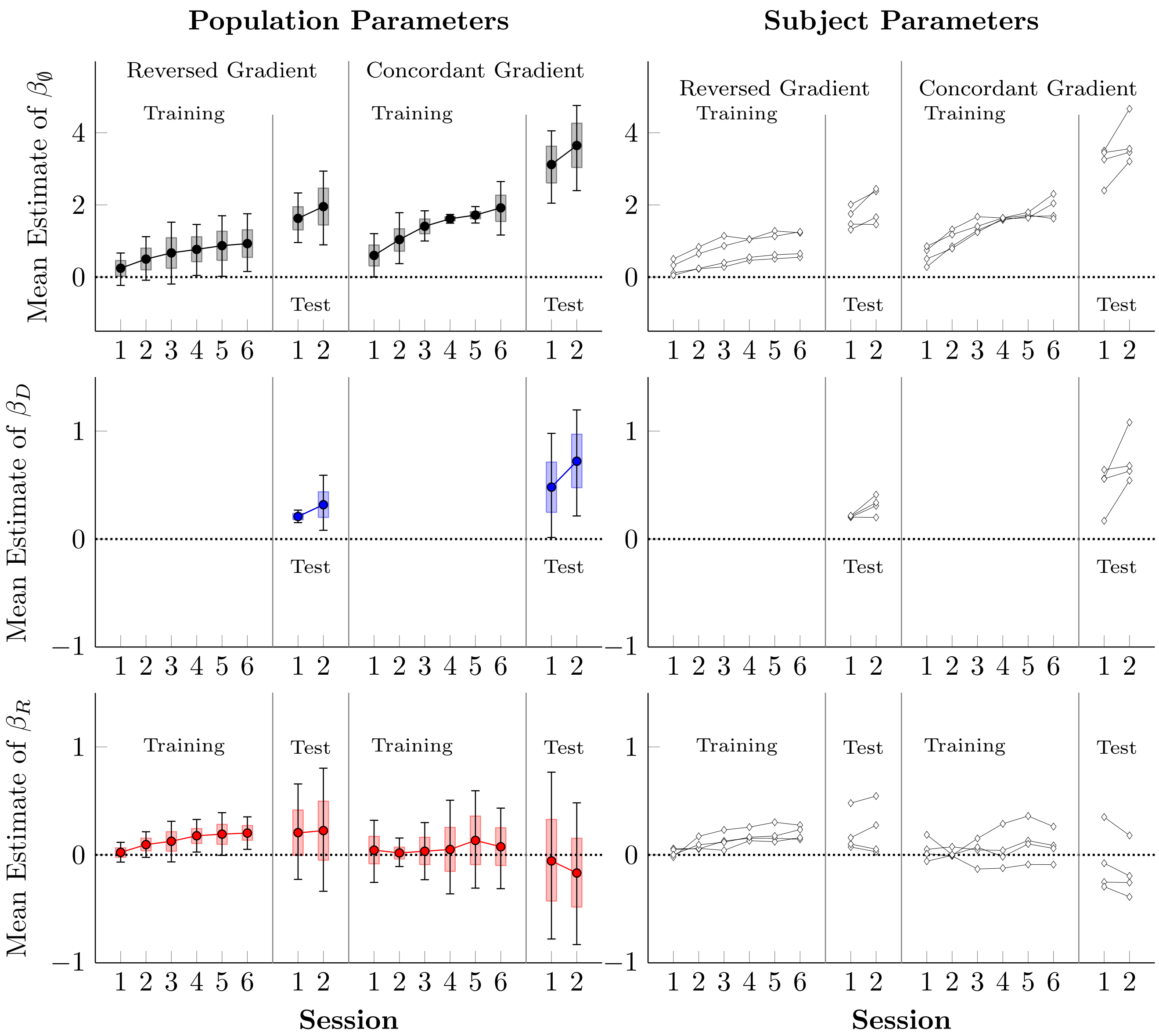


Figure 6. Mean population- and subject-level estimates of parameters in Experiment 2, presented on a logit scale. Filled circles correspond to population estimates, and white diamonds correspond to subject-level parameters for each subject. Whiskers correspond to the 99% credible interval, and boxes correspond to the 80% credible interval.

Gradient condition than in the Reversed condition. Additionally, although response accuracy was generally growing throughout training, so too was the β_R parameter in the Reversed Condition, suggesting a growing distortion to the response accuracies. No clear time course was evident among the values of β_R during training that used Concordant Gradients.

Figure 6 depicts average population-level parameters (left) and subject-level parameters (right) across Experiment 2. These are generally consistent with the impressions given by Figure 5: The growth of the intercept term $\beta_\emptyset$ during training appears a bit slower in the Reversed Gradient condition than in the Concordant Gradient condition. Nevertheless, both groups performed above chance and showed clearly positive symbolic distance effects at test, consistent with being able to perform TI in both cases. A larger distortion from the reward gradient was suggested in the Reversed Gradient condition, but the distribution

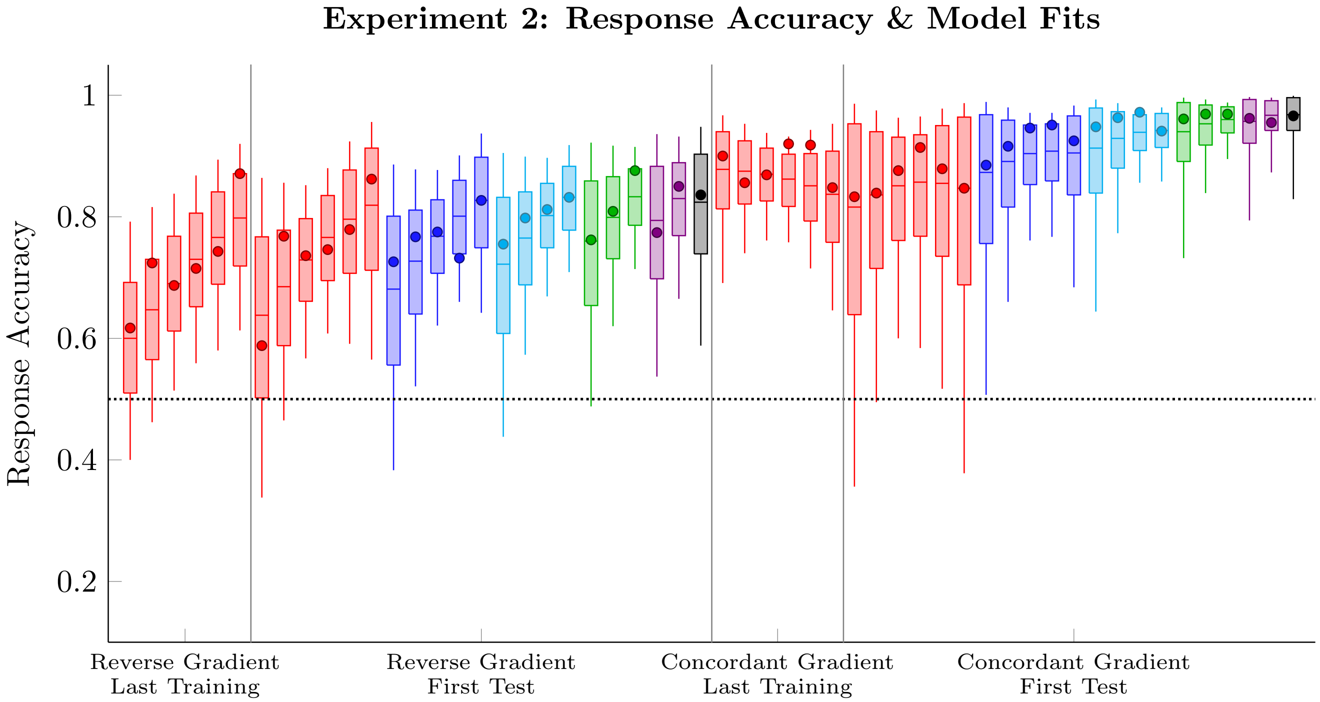


Figure 7. Response accuracy for each pair during the last phases of training and the first two phases of testing. Points depict the means subject response accuracies, as computed directly from the data. Box-and-whisker plots correspond to model estimates of response accuracy, with whiskers corresponding to the 99% credible interval and boxes corresponding to the 80% credible interval. Pairs are sorted by distance ($D = 1$ in red, $D = 2$ in blue, $D = 3$ in cyan, $D = 4$ in green, $D = 5$ in purple, and $D = 6$ in black,), then by rank (AB, BC, CD, etc.).

of these effects overlapped with zero in most phases.

Figure 7 realizes the implications for the logistic regression models, using only the parameters during the last session of training and the first session of testing. As in Experiment 1, the Reversed Gradient manipulation creates a distortion in performance that impairs early list pairs (e.g. AB, BC) but benefits late list pairs (e.g. EF, FG) during both training and testing. Despite this distortion, non-adjacent pairs, including all critical test pairs, appear to be above average at test, with a mild symbolic distance effect in evidence. Contrastingly, the Concordant Gradient condition does not appear to have distorted performance in the opposite direction. Accuracy was generally flat among adjacent pairs during training and minimal during testing.

Bayesian model of stimulus positions

In order to provide a computationally tractable Bayesian model of behavior, it was presumed that the position of each stimulus was represented by a normal distribution with parameters μ_i and σ_i for stimulus i. On each trial, a random value was drawn from each distribution, and whichever random draw was larger was the stimulus that was selected. When distributions overlapped, the distribution with the higher mean was more likely to be selected, but the alternative items was still chosen some amount of the time. This recapitulates the logic behind the betasort model (Jensen et al., 2015), but made use of normal distributions instead to facilitate parameter estimation. This logic is presented visually in Figure 8

Like the betasort model, there was also a parameter specifying the probability that subjects would disregard the representation and make a completely random response. This was included because monkeys in many experimental contexts never reach ceiling performance, instead maintaining some error

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018

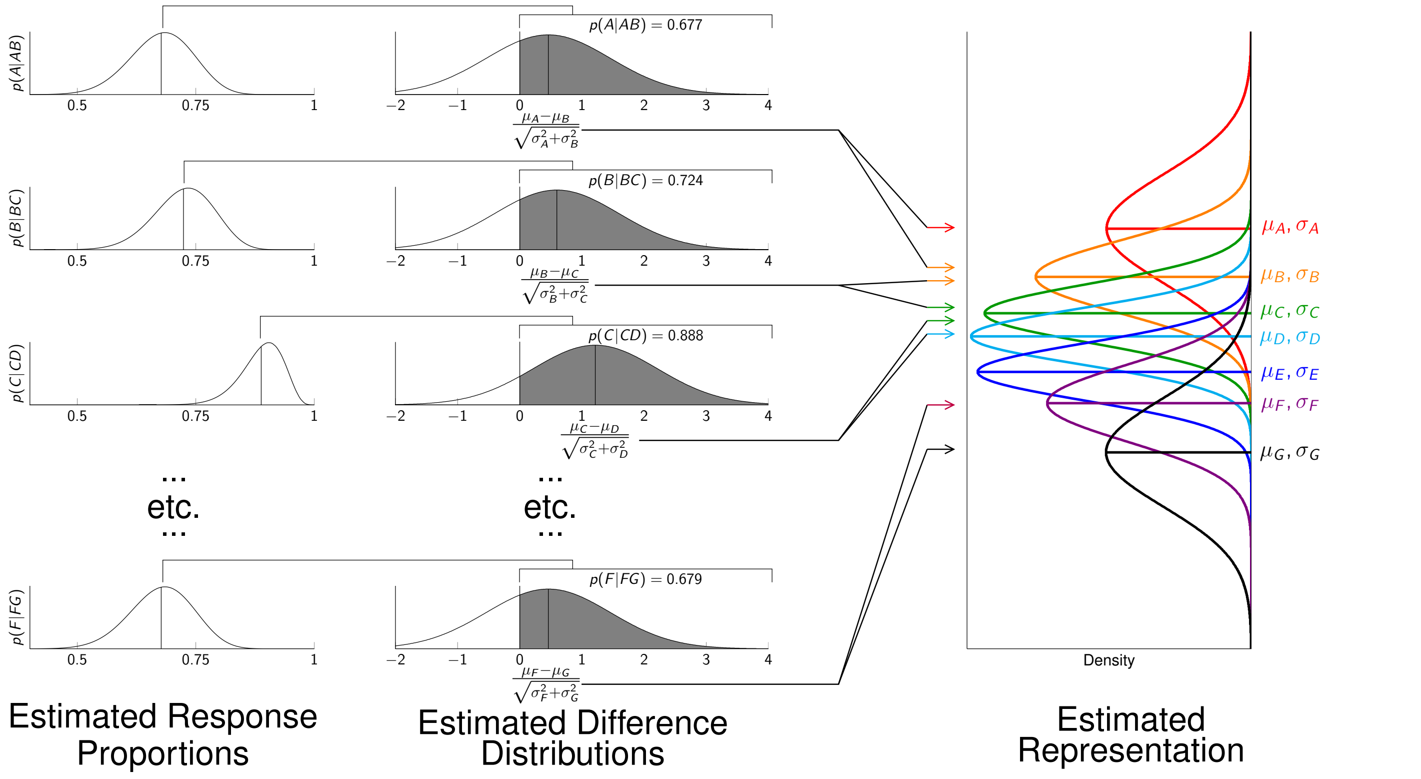


Figure 8. Model of list representation. Subjects are presumed to make use of a linear representation with uncertain stimulus positions. We assume this representation takes the form of a normal distribution with some mean and standard deviation for each stimulus. To infer these parameters, we estimated p("correct") for each pair and transformed this to the area above zero of some z distribution. Inferring the parameters in our representation is then done as a simultaneous estimation problem, implemented using Stan.

rate regardless of how much additional training they receive. This parameter is denoted by θ, where $0.0 < \theta < 1.0$.

The odds that one normal distribution yields a larger value than a second normal distribution is identical to the odds that the difference between the two values is positive. Since the variances of normal distributions are additive under subtraction, and since there was a probability of θ of an arbitrary response, the overall probability of a stimulus A in the pairing AB is given as follows:

$$p\left(A|AB\right) = \frac{\theta}{2} + (1-\theta)\int_0^\infty \mathcal{N}\left(x|\mu_A - \mu_B, \sqrt{\sigma_A^2 + \sigma_B^2}\right) dx \qquad (1)$$

Since there are seven stimuli, behavior is modeled in terms of 15 parameters: μ_i and σ_i for each stimulus, as well as θ. Estimating these requires solving a simultaneous estimation problem, where every pair of stimuli has its own version of Equation 1. The electronic supplement includes a Stan model that solves this problem as a multi-level model, yielding estimates of these parameters for both the population and for each subject. Note that because μ_i is unitless and defined only relative to the means of the other stimuli, values of μ_i and σ_i may be rescaled arbitrarily, so long as the scale is applied consistently for all position parameters. In order give a common scale to the positions for the purposes of plotting their estimates, the μ parameters were centered at zero, then all μ and σ parameters were divided by the sample standard deviation of the μ parameters. From a performance perspective, the model is unchanged under rescaling, since its comparisons are all relative between stimuli. Figure 9) plots the population means of

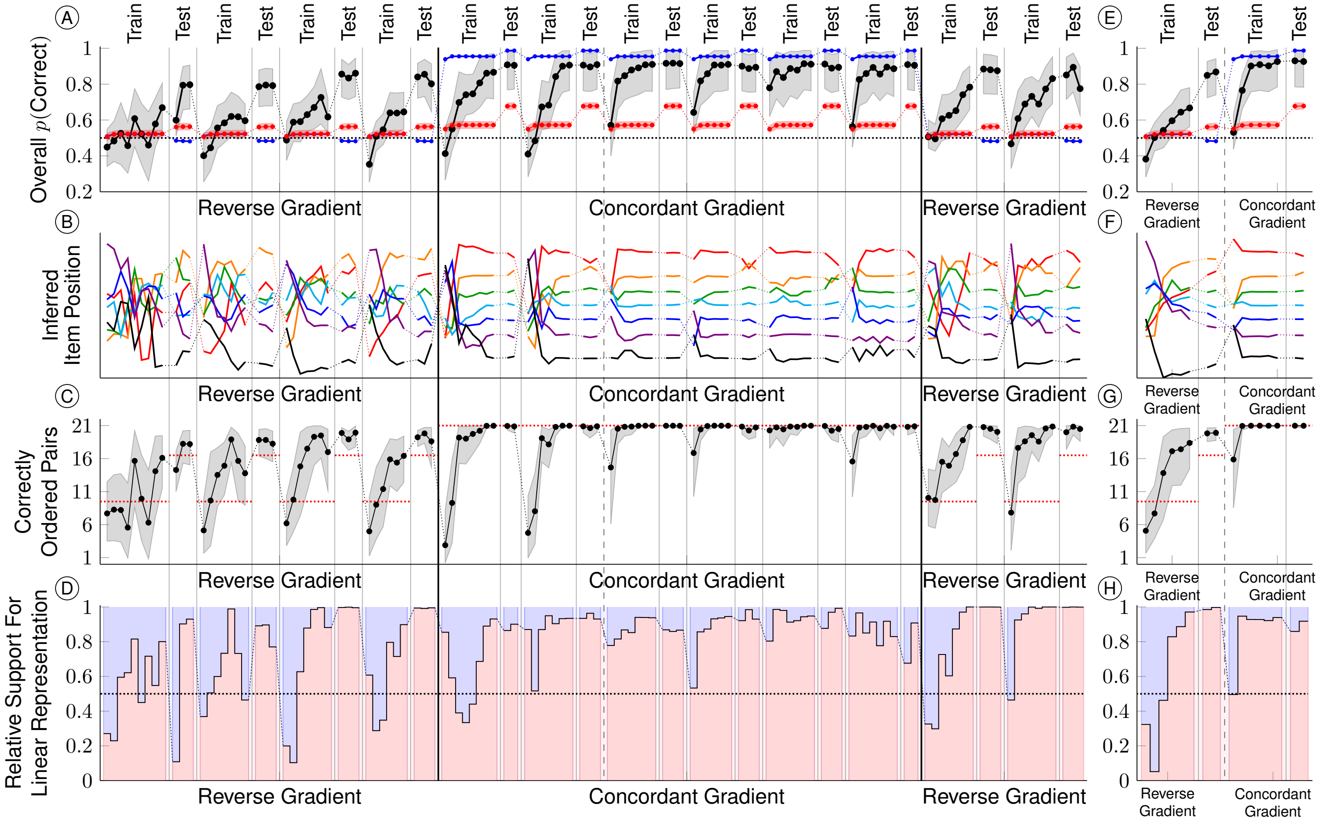


Figure 9. Bayesian estimates of representation of position and corresponding performance estimates. (A) Population estimates of response accuracy (black) for each session, based on the model. Chance is indicated by the dotted line. Red points correspond to the exploratory Q-learner, fit to the observed data. Blue points correspond to the exploitative Q-learner, based on a previous study (Jensen et al., 2015). (B) Peaks of inferred position distributions of each stimulus in subjects' representations. Red = A, orange = B, green = C, cyan = D, blue = E, violet = F, and black = G. Subjects reconstructed the stimulus order in the concordant gradient condition, and did so approximately in the reverse gradient condition, with the exception of stimulus A. (C) The average number of the 21 stimulus pairs that fell in the correct order, based on the model estimates in the above panel. The red dotted line indicates how many pairs would be ordered correctly if subjects used expected values as the basis for ordering. (D) Support of the evidence for the positions in the inferred representation being organized according a strictly linear representation (in red), relative to the expected values (in blue), according to a BIC analysis. Subjects tended to be equivocal, or to favor the expected value, during the first few sessions of training. However, late in training, and throughout the testing phase, the inferred representations more closely resembled a linear ordering of stimuli with uniform spacing. (E) to (H) Same as left panels, but based on pooling data across the six lists.

these rescaled position estimates (panels B and F), as well as the estimated performance based on those models (panels A and E).

Comparison to model-free learning algorithms

Rather than merely assert rhetorically that expected or experienced reward value is insufficient to solve the Reverse Gradient condition, we used a model-free Q-learning algorithm (Watkins and Dayan, 1989) that can only use its experienced estimate of overall reward value to perform the task. Performance of these algorithms is also plotted in Figure 9) (panels A and E).

Although Q-learning ordinarily factors the 'maximum possible reward on the next trial' into its updating, transitive inference tasks are scheduled in such a way that a subject's choice on trial t has no impact on which choice alternatives are available on trial $t+1$. Consequently, this 'projection into the future' cancels itself out, leaving only basic reward prediction error updating of the memory vector Q:

$$Q\,(\text{choice}) \leftarrow (1-\alpha)\,Q\,(\text{choice}) + \alpha \cdot \text{Reward} \tag{2}$$

That is, on each trial, the algorithm uses the reward delivery (including a value of 0.0 when no reward is delivered) to update its value of the item chosen. Items that are not chosen are not updated.

This describe the memory vector, but not the criterion by which choices are made. In our implementation, choices are made on the basis of the softmax function (Luce, 1959), with an exponential term β:

$$p\,(A|AB) = \frac{\exp\left(Q\,(A)^{\beta}\right)}{\exp\left(Q\,(A)^{\beta}\right) + \exp\left(Q\,(B)^{\beta}\right)} \tag{3}$$

When $\beta < 1.0$, the result is an algorithm whose behavior is more exploratory, because it is more willing to choose response option B when A has a larger expected value. At the extreme, when $\beta = 0.0$, subjects are equally likely to choose A and B. Contrastingly, if $\beta > 1.0$, the algorithm will behave in an exploitative manner, because it will be biased more strongly toward the largest expected value. As β tends toward large values, preference strongly tilts toward exclusive selection of the item with the highest expected reward value.

Using Stan, performance of each subject was fit separately using Equations 2 and 3. The best-fitting parameters were α values of between 0.086 and 0.225, while the best-fitting β parameters values were between 0.365 and 0.526.

Further validation of Bayesian model

Figure 9 presents two additional ways of validating that performance is not consistent with expected reward value: A count of the “number of correctly ordered pairs” (panels C and G) and an information criterion measure (panels D and H).

The estimate of the number of correctly ordered pairs depends on three details. First, the expected values of the stimuli during training are not the same as during testing, so the number of pairs an expected reward value comparison would get right changes. Secondly, despite training only presenting six pairs, the model is always *capable* of making inferences about all 21 pairs. Consequently, the estimates given an idea of how many of the 21 pairs *would* be correct if the next trial was the start of testing. Thirdly, because the parameter estimates were estimated using Stan, they took the form of chains of MCMC results, not parameterized distributions. Since these chains of estimated value capture the covariation of estimates, it is value to assess how many pairs are correctly estimated for each iteration of the chain, and to use the distribution of resulting values to obtain a credible interval for the estimate. Note that, in the event of ties in expected value (e.g. $A = \frac{1}{1}$ and $B = \frac{2}{2}$), the pair is given a value of 0.5 since it is expected to be chosen correctly half the time. Even though the nominal expected reward value comparisons, on average, identify the correct order of most pairs during the testing phase of the reverse gradient condition, subjects outperform even that higher bar.

To determine whether the ‘weight of the evidence’ better favored a linear representation or one based on the nominal expected value, linear models were fit using the nominal values of each model (linear on the one hand, or as expected from the expected value on the other) as predictors, and the posterior distributions of stimulus positions from the Bayesian model as the outcomes. Each regression yielded a BIC score. The relative support of the evidence for the linear model over the expected value model, on a scale of 0.0 to 1.0, is given as follows:

$$\text{Support for Linear Representation} = \frac{\exp\left(-BIC_{\text{linear}}\right)}{\exp\left(-BIC_{\text{linear}}\right) + \exp\left(-BIC_{\text{expected value}}\right)} \tag{4}$$

Since the output of the MCMC analysis was a chain of position estimates, BIC scores were calculated for each iteration of the chain. Consequently, when Equation 4 was calculated, the mean BIC scores across all values in the chain were used.

Discussion

Although the Reverse Gradient condition continued to impose a distorting effect to the response accuracy of individual pairs, the manipulation did not undermine the process of transitive inference overall. Indeed, after its distorting effect was taken into account, the results in the Reversed Gradient condition showed all the hallmarks of successful transitive inference: Both above-chance responding and a positive symbolic distance effect. Transfer from adjacent pairs to all pairs is completely consistent with the cognitive account of TI, which undermines any claim that performance at test can be explained entirely in terms of associative mechanisms.

In the Concurrent Gradient condition, performance during both training and testing were characterized by high levels of accuracy and positive distance effects. Subjects performed better in the Concordant Gradient condition than in the Reverse Gradient condition, but did so without a distortion to accuracy among pairs with the same symbolic distance. For example, we might expect AB accuracy to be high and FG accuracy to be low (the opposite of the pattern seen in the Reverse Gradient condition). Despite the differential rewards, accuracy among adjacent pairs remained flat, rather than being distorted in favor of early list pairs.

So does Experiment 2 do more to illuminate the contributions of associative learning vs. cognitive representation? Consider, again, the contributions of proportion correct and reward size, this time limited to training on the adjacent pairs:

Stimulus	Proportion rewarded	Reverse Gradient reward size	RG Expected value (proportion · reward)	Concordant Gradient reward size	CG Expected value (proportion · reward)
A	1/1	1 drop	1 drop	6 drops	6 drop
B	1/2	2 drops	1 drop	5 drops	2.5 drops
C	1/2	3 drops	1.5 drops	4 drops	2 drops
D	1/2	4 drops	2 drops	3 drops	1.5 drops
E	1/2	5 drops	2.5 drops	2 drops	1 drop
F	1/2	6 drops	3 drop	1 drop	0.5 drops
G	0/1	0 drops	0 drops	0 drops	0 drops

During adjacent-pair training in the Reverse Gradient condition, the associative prediction of performance that is based on expected value is not monotonic. The pair FG should be above chance (1 drop vs. 0 drops), the pair AB should be at chance (1 drop vs. 1 drop), and the pairs BC, CD, DE, and EF should all be below chance (e.g. 1 drop vs. 1.5 drops for BC; or 2 drops vs. 2.5 drops for DE). If larger discrepancies in expected value should be easier to discriminate, BC should yield the lowest performance of all adjacent pairs. The pattern we observe during Reverse Gradient training does not resemble this pattern at all. Instead, all pairs appear above chance, and their ordering appears pretty close to monotonic (if anything, BC somewhat overperforms).

A similar mismatch between the associative prediction and the pattern of training performance is observed in the Concordant phase. The discrepancy between 6 drops and 2.5 drops in the case of AB is enormous, whereas the discrepancy between 2.5 drops and 2 drops for the pair BC is small. Despite this, both pairs yield similar accuracy during training. This sort of discrepancy contributes to the inability of the Q-learning algorithms to master the task (as in Figure 9), whereas a cognitive model of stimulus position, like the one we describe (also presented in Figure 9), is handily able to account for performance.

The case that expected value explanations fail to explain TI in designs that do not reward stimuli equally has been made previously (Lazareva and Wasserman, 2012; Jensen et al., 2015, 2017), but is especially clear in the Reversed Gradient condition because, as noted in the table above, that manipulation's effect on expected value should result in performance below chance for a majority of stimulus pairs.

When considering both performance during training and the evidence for TI during testing, a strictly associative account of performance appears incompatible with our observed results. The influence of some associative mechanism remains, evidenced in the distortion to performance in the Reversed Gradient condition. Nevertheless, some manner of model-based inference, such as a cognitive representation of the serial order, is needed to explain performance.

Interpreting the Q-learning comparisons

It is noteworthy that, of the two Q-learning algorithms, it was the exploratory models that appeared to most closely resembles the performance of subjects is one that is highly exploratory. This should be interpreted with a grain of salt, however, because the ”exploratory” algorithm's performance did not resemble that of subjects, as reported in the main manuscript. In practice, subjects often chose a response alternative that was the “wrong” choice, according to the values of Q at that trial. Consequently, only a relatively low value of α, which would permit these “errors,” could be an acceptable parameter.

As a result of these erroneous responses, the exploratory, best-fitting parameters were able to exceed chance performance in simulations of the task, both during training (by a hair) and testing (by a small amount). However, this above-chance performance was still far below what subjects were capable of, as plotted in Figure 2B and 2F of the main manuscript. The algorithm was also unable to fully capitalize on the beneficial reward information in the Concordant Gradient condition, making many errors during both training and testing.

The inclusion of both the α and β parameters gives Q-learning flexibility, so even though the exploratory parameters were the best-fitting, they were not necessarily the best that the algorithm could do in terms of total earnings. Consequently, we also implemented an exploitative algorithms ($\alpha = 0.15$, $\beta = 3.0$). This algorithms was very slightly above chance during Reverse Gradient training sessions and very slightly below chance during Reverse Gradient testing sessions, leading to a slightly lower return in those cases than the exploratory algorithm. However, during the Concordant Gradient phases, the exploitative algorithm performed near-perfectly, and these added rewards more than made up for chance performance during the Reverse Gradient sessions.

The important take-away of these simulations is two-fold. On the one hand, neither algorithm is able to explain performance under the Reverse Gradient condition. Although the best-fitting parameters allow Q-learning to exceed chance, they do so by allowing the algorithm to make frequent choices against its better judgment. On the other hand, the very thing that makes the exploratory parameters effective in the Reverse Gradient case (frequent “errors”) then puts a ceiling on how well it can perform under the Concordant Gradient condition. Meanwhile, the exploitation algorithm performs near-perfectly under the Concordant Gradient condition, but this is unsurprising because that condition can be solved by multiple ways. Thus, even if Q-learning were to dynamically adjust its β values from one session to the next, it still would not have the flexibility to perform as well as subjects.

GENERAL DISCUSSION

In Experiment 1, we presented subjects with all possible pairs of images from a 7-item ordered list of photographic stimuli, rewarding the selection of the item with an earlier rank. In some cases, subjects were rewarded uniformly for correct responses, and performance was consistently high under these circumstances. However, when a “reversed gradient” of reward magnitudes was first introduced (such that higher-ranked items, such as F, earned larger rewards despite being correct less often), performance began to deteriorate as the manipulation appeared to devalue early list items relative to late list items. This result was closely consistent with the predictions of an associative model. However, when this manipulation was introduced a second time in the final phase of the experiment, its distorting effect was less severe, and did not prevent subjects from choosing correct responses most of the time for all stimulus pairs.

To further illuminate the mechanisms underlying this behavior, Experiment 2 trained novel 7-item lists using the adjacent-pair-training/all-pair-testing paradigm. Additionally, the Reversed Gradient condition was complemented by a second distribution of reward magnitudes, the Concordant Gradient condition. As in Experiment 1, the Reversed Gradient condition continues to exert an influence on the relative accuracy of stimulus pairs during both training and testing, but contrary to the expectations of a model based solely on associative mechanisms, accuracy remained above chance in general and displayed all the hallmarks of transitive inference at test (including a clear symbolic distance effect). Meanwhile, although the Concordant Gradient condition displayed higher performance overall, the differential rewards had no clear distorting effect on the relative accuracy of the individual pairs, either during training or testing.

Collectively, these results suggest that, on the one hand, the expected value of reward (in the model-free sense) is being calculated by subjects, and plays some role in action selection. On the other hand, however, this study provides strong evidence that such associative mechanisms cannot be the sole factor explaining performance, both because of the improved performance in Phase 4 of Experiment 1 and the pattern of results consistent with transitive inferences during testing in Experiment 2. Although studies of TI often frame performance as arising either from associative mechanisms or from cognitive representations, our results are best explained by some persistent joint influence between two systems acting in parallel.

A substantial and growing body of work in neuroscience supports the view of value systems that are at once complementary and dissociable. Popular targets for these two systems are the lateral prefrontal cortex and the striatum (reviewed by Tanaka et al., 2015). Although the contributions of these areas are framed in terms of “model-based” vs. “model-free” inference respectively (e.g. Ito and Doya, 2011), we prefer a less vague and more concrete nomenclature. Striatal circuits appear to be strictly retrospective and experience-driven (Daw et al., 2005), whereas prefrontal circuits demonstrate an ability to prospectively infer the values of novel stimuli (Pan et al., 2014).

Although these advances in our understanding of brain circuitry help to move us past the either/or logic of the cognitive revolution, it is nevertheless also the case that our experimental results, and others, cannot be interpreted as merely passive averaging of two systems. It appears as though associative mechanisms interact with abstract inference in some cases, such as the distortion of pairwise accuracy we observed in the Reversed Gradient condition, but associative mechanisms are at other times seemingly ignored entirely, as in the case of the undistorted inferences in the Concordant Gradient condition. Similarly, associative models predict that when some stimulus pairs are presented much more often than others, a substantial distortion in subsequent transitive inference should be observed. However, despite evidence that such associative signals are likely calculated, studies that included massed presentations of certain stimulus pairs yield no evidence at all of distorted inference (Lazareva and Wasserman, 2012; Jensen et al., 2017). While it should no longer be controversial that the brain makes both cognitive inferences and expected value calculations, the question remains of how these two calculations interact, and the conditions under which one supersedes the other in determining behavior.

ACKNOWLEDGMENTS

We thank David Freshwater, Aliza Gross, Grant Spencer, and Natalie Tanner for assistance with data collection.

AUTHOR CONTRIBUTIONS

GJ, VPF, and HST conceived the experiments. GJ and YA acquired data. GJ wrote the task software and performed analyses. GJ, YA, VPF, and HST wrote the paper.

REFERENCES

Allen, C. (2006). Transitive inference in animals: Reasoning or conditioned associations? In Hurley, S. and Nudds, M., editors, *Rational Animals*, pages 175–186. Oxford University Press, Oxford, UK.

Berdyyeva, T. K. and Olson, C. R. (2011). Relation of ordinal position signals to the expectation of reward and passage of time in four areas of the macaque frontal cortex. *Journal of Neurophysiology*, 105:2547–2559.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32.

Daw, N., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711.

Epstein, R. A., Patai, E. Z., Julian, J. B., and Spiers, H. J. (2017). The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, 20:1504–1513.

Gazes, R. P., Chee, N. W., and Hampton, R. R. (2017). Cognitive mechanisms for transitive inference performance in rhesus monkeys: Measuring the influence of associative strength and inferred order. *Journal of Experimental Psychology: Animal Behavior Processes*, 38:331–345.

Ito, M. and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21:368–373.

Jensen, G. (2017). Serial learning. In Call, J., Burghardt, G. M., Pepperberg, I. M., Snowdon, C. T., and Zentall, T., editors, *APA Handbook of Comparative Psychology: Perception, Learning, and Cognition*, pages 385–409. American Psychological Association.

Jensen, G., Alkan, Y., Muñoz, F., Ferrera, V. P., and Terrace, H. S. (2017). Transitive inference in humans (*Homo sapiens*) and rhesus macaques (*Macaca mulatta*) after massed training of the last two list items. *Journal of Comparative Psychology*, 131:231–245.

Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P., and Terrace, H. S. (2015). Implicit value updating explains transitive inference performance: The betasort model. *PLOS Computational Biology*, 11:e1004523.

Lazareva, O. F. and Wasserman, E. A. (2012). Transitive inference in pigeons: Measuring the associative value of stimuli b and d. *Behavioural Processes*, 89:244–255.

Luce, D. (1959). *Individual Choice Behavior*. Wiley, New York, NY.

McGonigle, B. O. and Chalmers, M. (1977). Are monkeys logical? *Nature*, 267:694–696.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., and et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–533.

Pan, X., Fan, H., Sawa, K., Tsuda, I., Tsukada, M., and Sakagami, M. (2014). Reward inference by primate prefrontal and striatal neurons. *Journal of Neuroscience*, 34:1380–1396.

Pan, X., Sawa, K., Tsuda, I., Tsukada, M., and Sakagami, M. (2008). Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. *Nature Neuroscience*, 11:703–712.

Tanaka, S., Pan, X., Oguchi, M., Taylor, J. E., and Sakagami, M. (2015). Dissociable functions of reward inference in the lateral prefrontal cortex and the striatum. *Frontiers in Psychology*, 6:Article 995.

Tanner, N., Jensen, G., Ferrera, V. P., and Terrace, H. S. (2015). Inferential learning of serial order of perceptual categories by rhesus monkeys (*Macaca mulatta*). *Journal of Neuroscience*, 37:6268–6276.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55:189–208.

Vasconcelos, M. (2008). Transitive inference in non-human animals: An empirical and theoretical analysis. *Behavioural Processes*, 78:313–334.

von Fersen, L., Wynne, C. D. L., Delius, J. D., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17:334–341.

Watkins, C. J. C. H. and Dayan, P. (1989). Q-learning. *Machine Learning*, 8:279–292.

Wynne, C. D. L. (1995). Reinforcement accounts for transitive inference performance. *Animal Learning & Behavior*, 23:207–217.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26889v3 | CC BY 4.0 Open Access | rec: 27 Oct 2018, publ: 27 Oct 2018