Distributed representations for human inference

Zhenglong Zhou*, Dhairyya Singh, Marlie C. Tandoc, & Anna C. Schapiro*
Department of Psychology, University of Pennsylvania

* Authors for correspondence: zzhou34@sas.upenn.edu, aschapir@sas.upenn.edu

Abstract

Neural representations can be characterized as falling along a continuum, from distributed representations, in which neurons are responsive to many related features of the environment, to localist representations, where neurons orthogonalize activity patterns despite any input similarity. Distributed representations support powerful learning in neural network models and have been posited to exist throughout the brain, but it is unclear under what conditions humans acquire these representations and what computational advantages they may confer. In a series of behavioral experiments, we present evidence that interleaved exposure to new information facilitates the rapid formation of distributed representations in humans. As in neural network models with distributed representations, interleaved learning supports fast and automatic recognition of item relatedness, affords efficient generalization, and is especially critical for inference when learning requires statistical integration of noisy information over time. We use the data to adjudicate between several existing computational models of human memory and inference. The results demonstrate the power of interleaved learning and implicate the use of distributed representations in human inference.

Introduction

How does the brain encode new information? What is the nature of the representations that it builds? One of the most fundamental dimensions of these questions, and a subject of much debate across areas of neuroscience, is representational overlap—the degree to which neurons code for multiple entities. At one end of the spectrum, information can be encoded using *distributed representations* (Hinton, 1984), in which populations of neurons are responsive to many related features of the environment. On the other end, information can be orthogonalized in *localist representations* (Doumas & Hummel 2005), with individual neurons coding for few entities. We know that these types of representations have markedly different properties in computational models. Distributed representations are extremely good at representing structure in data, empowering their success in many challenging domains (Hassabis et al., 2017). But these representations are famously susceptible to interference (McCloskey & Cohen 1989; Hadsell et al., 2020). Localist representations are robust to interference but less powerful in large scale applications, and they require regularities across experiences to be computed with some effort at retrieval (Kumaran & McClelland, 2012).

The brain likely makes use of both types of representation. Neural evidence suggests that distributed representations, some of which resemble those found in neural network models, exist throughout the brain, especially in the neocortex (Yamins et al., 2014; McClelland et al., 1995; Desimone et al., 1984). There is also evidence for localist (i.e. *pattern-separated*) representations, primarily in the hippocampus, where they may serve to avoid interference between episodes in memory (O'Keefe 1976; Norman & O'Reilly 2003). Despite some evidence of their presence in different brain structures, it remains unclear when and how we acquire these two different forms of representation, and what the consequences of the formation of each might be for adaptive human behavior.

A fundamental property of neural network models employing distributed representations is that effective behavior requires exposure to entities in interleaved order (McClelland et al., 1995). Interleaving allows the learning algorithm to carefully construct representations that reflect shared structure across inputs. By contrast, learning in a blocked fashion, where one set of information is encountered before a second set, tends to overwrite knowledge of the first set, to the extent that the sets are related (i.e., 'catastrophic interference'; McCloskey & Cohen 1989; McClelland et al., 2020). We hypothesized that interleaved exposure to information would similarly facilitate the formation of distributed representations in the human brain. One barrier to testing this idea is that neocortex, where distributed representations are typically thought to be housed, likely learns complex novel information slowly, on the order of days or weeks (Recanzone et al., 1993; Schoups et al., 2001). However, we recently developed a neural network model (Schapiro et al., 2017) which instantiates the hypothesis that a particular circuit within the hippocampus—the monosynaptic pathway connecting entorhinal cortex to subfield CA1—employs distributed representations that are responsible for the ability of the hippocampus to detect regularities across experiences on the

timescale of minutes (Schapiro et al., 2012). We will refer to the model here as C-HORSE: Complementary Hippocampal Operations for Representing Statistics and Episodes. Though the hippocampus is known for its pattern-separated representations, rodent work indicates that these reside mainly in the DG and CA3 subfields, while representations appear more overlapping in CA1 (Leutgeb et al., 2004). Our model of the hippocampus thus motivates searching for the formation of distributed representations across the timescale of one human experimental session, in the context of a hippocampally-dependent task.

We adopted the hippocampally-dependent associative inference task (Bunsey & Eichenbaum, 1996; Preston et al., 2004), in which participants learn novel object pairs AB and BC and are then tested on the unobserved indirect AC relationship. The basic associative inference task can be solved using either a distributed or localist strategy. The distributed strategy, which is implemented in hippocampal subfield CA1 in C-HORSE (Schapiro et al. 2017), learns overlapping representations of A and C by virtue of both being paired with B, and thus supports a direct and immediate association between A and C at test. The localist strategy, proposed by Kumaran & McClelland (2012) in the REMERGE model, posits that the hippocampus relies on recirculating activity at retrieval across non-overlapping representations of AB and BC pairs (localist representations of each learning episode) in subfields DG and CA3. C-HORSE proposes that distributed and localist representations are both present in the hippocampus, housed in these different subfields (Schapiro et al., 2017). We thus expect that both strategies are available to the system during associative inference. However, these two strategies differ in a few crucial ways. First, the distributed strategy promotes implicit, automatic association between A and C at test, whereas the localist strategy requires additional, and perhaps more explicit, computation to achieve successful inference. Second, distributed representations should more readily promote generalization of novel information: Acquiring new knowledge about A should immediately transfer to C by virtue of their shared neural substrate. Lastly, distributed representations should exhibit more robust graded sensitivity to statistical regularities (Rogers & McClelland 2004). Each participant in our experiments learned some sets of AB and BC pairs in interleaved order, and other sets in blocked order. If humans learn representations similarly to distributed neural network models, only interleaved exposure should result in behavior that shows these properties. We provide evidence that, indeed, interleaved exposure results in these behaviors. The results are consistent with simulations from C-HORSE, which includes both distributed and localist representations, but not with REMERGE, containing only localist representations, or the Temporal Context Model (Howard & Kahana 2002; Howard et al., 2009), which relies on distributed representations. Together, the results reveal the benefits of interleaved learning and suggest that humans rapidly form distributed representations, in addition to more separated representations, in the service of efficient inference.

Results

Interleaved learning drives recognition of indirect relatedness. In the first three experiments (Exps 1a-c), during learning each participant learned triads of direct AB and BC associations, with the direct pairs for some triads presented in interleaved order and others in blocked order (Fig. 1a). During the presentation of each direct pair, they were told to visualize a story linking the objects. After learning, participants completed two tasks that probed their memory for object associations: a speeded recognition task, requiring relatively rapid, implicit judgments (Fig. 1b) followed by an explicit inference task, in which participants more slowly and deliberately identified indirect associations (Fig. 1c).

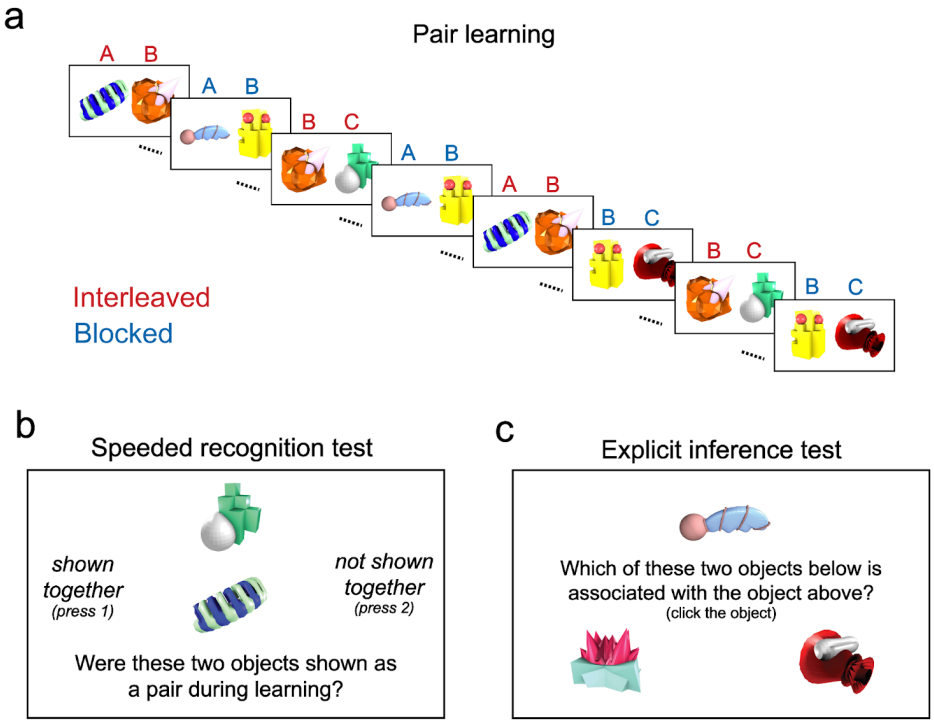

Figure 1. Exps 1a-c design. (**a**) During learning, participants viewed pairs of objects. Pairs that shared a common object (i.e., AB and BC) appeared either in an interleaved (red) or blocked (blue) order. (**b**) In the speeded recognition test, participants quickly judged whether two objects were paired during learning. (**c**) In the explicit inference test, participants selected which of two objects was indirectly related to a cue object.

In the speeded recognition task, on each trial, participants saw a pair of objects for 1500 ms and had 3500 ms total to respond as to whether they had seen that pair during learning (Fig. 1b). For all pairs of objects except those forming an AB or BC pair, the correct answer was to indicate that they were *not* shown together during learning. This decision could be challenging for the indirectly related AC pairs, as they were associated but not directly studied together. We predicted that this would be especially challenging in the interleaved condition. If interleaving leads to overlapping

representations, there would be more difficulty rejecting ACs, which would manifest in increased response time and/or false alarms. For the blocked condition, we predicted a reliance on localist representations of AB and BC, which should prevent this kind of confusion on AC recognition trials.

For the subsequent inference task, participants were explicitly instructed that two objects form an indirect AC association if they were both paired with the same object. On each trial (Fig. 1c), participants were then given 7000 ms to identify the object that was indirectly associated with a cue object against a matched foil (see Methods). In this explicit setting where participants are encouraged to search for indirect links amongst objects and have enough time to do so, we predicted no difference between AC inference in the two conditions, as localist and distributed strategies should both be available to solve this task.

In Exp 1a, consistent with our predictions, in the speeded recognition task participants were slower to correctly reject ACs as not having been paired during learning in the interleaved relative to blocked condition (Fig. 2a): On correct trials, participants responded slower to interleaved ACs than to blocked ACs ($t(25)=2.46$, $p=0.02$; Fig. 2a) and to interleaved foils (i.e., pairs of objects from different interleaved triads; $t(25)=5.67$, $p<0.001$; Fig. S1), whereas reaction time (RT) did not significantly differ between blocked ACs and foils ($t(25)=0.96$, $p=0.34$; Fig. S1) nor between interleaved and blocked foils ($t(25)=-1.78$, $p=0.09$; Fig. S2). RT did not differ between interleaved and blocked direct pairs ($t(25)=-1.40$, $p=0.17$; Fig. S10). Participants also exhibited higher false alarm rates to interleaved than to blocked ACs ($t(25)=2.76$, $p=0.01$; Fig. 2b), meaning they were more likely to incorrectly indicate that they had seen these indirect pairs when the direct constituents were presented in interleaved order. There was no difference in false alarm rates between interleaved and blocked foils ($t(25)=0.50$, $p=0.62$; Fig. S3). These results support our prediction of an advantage of interleaved exposure for forming representations that support direct and automatic AC association.

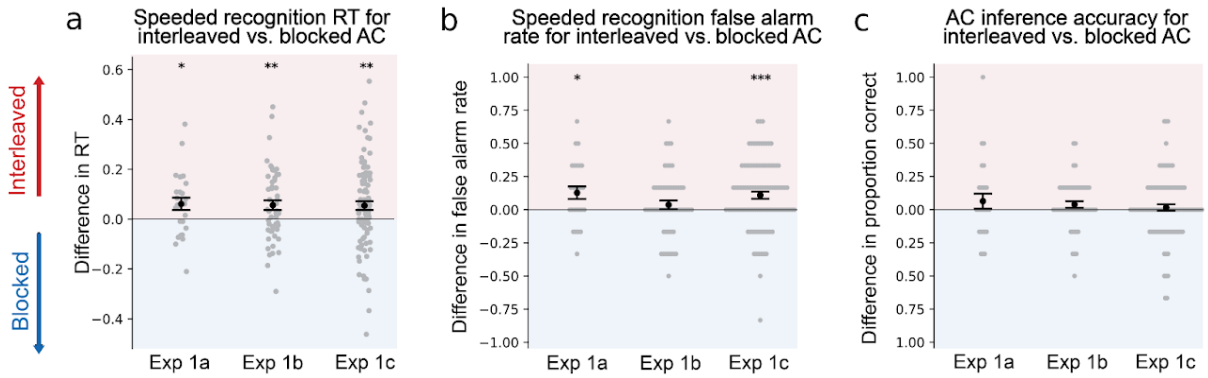

Figure 2. Exps 1a-c results. (**a**) RTs for interleaved vs. blocked AC trials during speeded recognition. (**b**) False alarm rates for interleaved vs. blocked AC trials during speeded recognition. (**c**) Accuracy differences

between interleaved and blocked AC trials during explicit inference. Error bars represent +/- 1 SEM. *p<0.05; **p<0.01; ***p<0.001.

To increase confidence in the results from Exp 1a, we preregistered (https://osf.io/ag42z/) and ran a replication of the experiment as Exp 1b. In Exp 1b, participants again responded significantly more slowly to interleaved ACs than to blocked ACs (t(52)=2.82, p=0.007; Fig. 2a) and to interleaved foils (t(52)=3.84, p<0.001; Fig. S1), and RT was not significantly different between blocked ACs and blocked foils (t(52)=1.64, p=0.11; Fig. S1) or between interleaved and blocked foils (t(52)=0.24, p=0.81; Fig. S2). There was again no difference in RT between interleaved and blocked direct pairs (t(52)=-1.61, p=0.11; Fig. S11). Though numerically in the same direction, Exp 1b did not replicate the significant difference in false alarm rates between interleaved and blocked AC trials (t(52)=1.14, p=0.26; Fig. 2b).

Exps 1a-b measured participants' memory for AB and BC pairs only through the speeded recognition task, which does not assess the ability to explicitly identify direct associations against competing alternatives. To verify that participants can do this equally well for the two conditions, we performed an additional replication study, Exp 1c, which additionally assessed explicit identification of AB/BC pairs against foils at the end of the experiment. Results of Exp 1c again demonstrated stronger AC associations in the interleaved condition: During the speeded recognition task, participants responded more slowly to interleaved ACs than to blocked ACs (t(94)=3.00, p=0.003; Fig. 2a) and to matched foils (t(94)=5.22, p<0.001; Fig. S1). RT again did not significantly differ between interleaved and blocked direct pairs (t(94)=0.35, p=0.73; Fig. S12). Here we again observed, as in Exp 1a, higher false alarm rates to interleaved than to blocked ACs (t(94)=4.55, p<0.001; Fig. 2b). RT (t(96)=1.69, p=0.09; Fig. S2) and false alarm rate (t(96)=1.20, p=0.23; Fig. S3) were not significantly different between interleaved and blocked foils. In the task that explicitly assessed AB/BC memory, participants showed matched memory between the two conditions (t(99)=-0.36, p=0.72; Fig. S4a) and between AB and BC pairs in both conditions (interleaved: t(99)=-0.34, p=0.73; Fig. S15; blocked: t(99)=-1.23, p=0.22). This result combined with the lack of differences in direct pair RTs suggest that the observed advantage for rapid recognition of interleaved AC associations is unlikely to be due to a difference in direct pair memory between conditions. Participants responded slower during the explicit inference task than in this explicit direct pair assessment in both the interleaved (t(99)=17.47, p<0.001; Fig. S9) and the blocked condition (t(98)=15.07, p<0.001). We also observed slower responses to indirect than to direct speeded recognition trials across Exps 1a (interleaved: t(25)=7.88, p<0.001; blocked: t(25)=5.28, p<0.001; Fig. S8), 1b (interleaved: t(52)=11.11, p<0.001; blocked: t(25)=5.74, p<0.001), and 1c (interleaved: t(95)=9.56, p<0.001; blocked: t(96)=8.51, p<0.001), suggesting that direct associations were always more accessible than indirect associations. Exps 1a, 1b, and 1c together provide strong evidence that interleaved exposure facilitates direct and automatic association of indirectly related items.

No differences between blocked and interleaved exposure in an explicit assessment of inference. As predicted, across Exps 1a, 1b, and 1c, we did not observe significant accuracy differences between inference performance for interleaved and blocked ACs in the explicit inference task (1a: $t(25)=1.12$, $p=0.27$; 1b: $t(53)=1.60$, $p=0.12$; 1c: $t(99)=0.71$, $p=0.48$; Fig. 2c). There were also no differences in RT (1a: $t(24)=-0.46$, $p=0.65$, Fig. S10; 1b: $t(53)=-0.05$, $p=0.96$, Fig. S11; 1c: $t(98)=-1.26$, $p=0.21$, Fig. S12). Equivalent performance between the two conditions in this relatively slow and explicit setting indicates that participants were able to successfully link blocked AC associations despite displaying little sensitivity to these associations in the implicit speeded setting. This pattern of results is consistent with the use of relatively slower recurrent computation across localist representations, as hypothesized by REMERGE (Kumaran & McClelland 2012; Tamminen et al., 2015).

Interleaved exposure affords rapid generalization. Exps 1a-c demonstrated an enhanced sensitivity to indirect associations after interleaved learning that could be viewed as disadvantageous, as the task asked participants to report only the directly learned associations. In the next set of experiments, we aimed to highlight situations in which representations learned through interleaving should be beneficial. In neural network models, distributed representations support automatic generalization (Hinton, 1984): learning new knowledge about an object automatically allows that knowledge to be transferred to related objects by virtue of the overlap in their representations. By contrast, localist codes do not as naturally support such transfer (they can, but only with additional carefully constructed recurrent computation; McClelland et al., 1986)

In Exp 2, we tested whether behavior after interleaved learning reflects this quality of distributed representations. After participants learned interleaved and blocked AB and BC associations as in Exp 1 (Fig. 3a), they then learned to associate novel nonsense properties with a subset of the objects (Fig. 3b). For each ABC triad, participants associated a unique property with either A or C. Participants then completed three tests: property memory, property generalization, and explicit direct pair recognition. In the property memory test (Fig. 3c), we assessed participants' memory for the trained property-object pairings, with no expectation of a difference in performance between conditions. In the property generalization test, we assessed their ability to generalize the novel properties to indirectly related objects (Fig. 3d). We did not explicitly instruct them about the presence of indirectly related objects, as in the explicit inference task above. Participants had 4000 ms to view the options and make a decision. We predicted superior property generalization performance for interleaved relative to blocked ACs. Finally, participants were assessed on their memory of AB and BC pairings as in Exp 1c (not shown in Fig. 3).

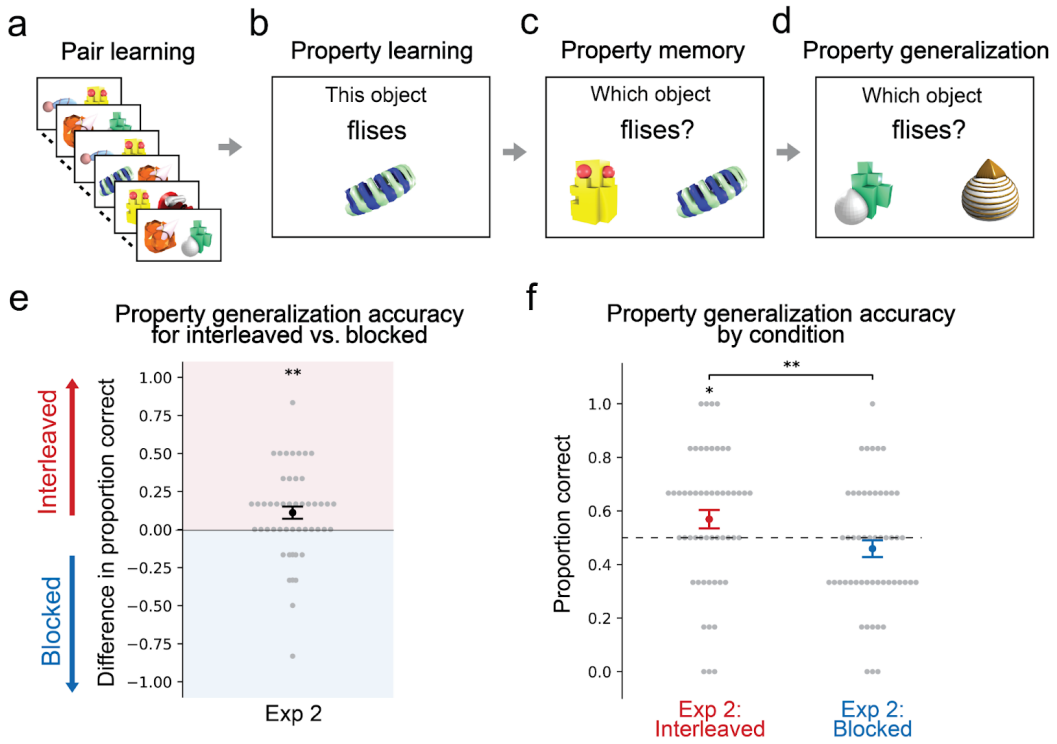

Figure 3. Exp 2 design and results. (a) Participants viewed ABC triads in either interleaved or a blocked order as in Exp 1. (b) During property learning, participants learned to associate a subset of objects with novel nonsense properties. (c) The property memory test assessed participants' memory for these trained pairings. (d) In the property generalization task, participants were asked to identify which of two objects was more likely to have the novel property, one of which was transitively related to an object that participants had learned had that property. (e) Difference in property generalization accuracy between interleaved and blocked triads. (f) Property generalization accuracy broken down by condition.

As expected, participants' memory for trained object-property associations was not different between interleaved and blocked items in the property memory test (t(52)=-0.00, p=1.00; Fig. S6). However, in the property generalization task, participants performed significantly better for interleaved than blocked triads (t(52)=2.80, p=0.007; Fig. 3e). Unlike Exp 1c, participants here showed superior explicit memory for interleaved relative to blocked direct pairs (t(52)=2.09, p=0.04; Fig. S5a). This raises the possibility that weak property generalization performance in the blocked condition could be due to poorer memory of blocked AB/BC pairs, though there was no reliable correlation (r=0.021, p=0.83). To rule out this possibility, we ran analyses restricted to a subsample of subjects (n=44) who had matched performance between conditions on direct pairs (t(43)=0.12, p=0.91), by excluding subjects who showed the weakest memory for blocked relative to interleaved pairs. Analyses based on this sample again revealed matched memory for object-property associations between conditions (t(43)=0.16, p=0.87) and superior property generalization performance for interleaved relative to blocked triads (t(43)=2.37, p=0.02). Thus, the advantage of

interleaved associations during property generalization is unlikely to be due to lack of retention of direct pairs in the blocked condition. In sum, Exp 2 provides evidence that representations formed via interleaved exposure afford automatic generalization of new knowledge to indirectly related objects.

Only interleaved exposure permits successful inference when learning requires statistical integration over time. Thus far, we have demonstrated that interleaved learning influences behavior in tasks that require relatively rapid judgments. Exps 1a-c did not identify a difference between interleaved and blocked conditions in the slower, explicit AC inference task. Indeed, we did not predict such a difference, due to the availability of an alternative strategy in this case: indirect association through recurrent computation across localist representations, as proposed by the REMERGE model (Kumaran & McClelland, 2012). Are there certain learning problems where, even given sufficient opportunity for recurrent processing, only interleaved exposure can support successful behavior?

One situation in which we expect the localist strategy to fail is when direct associations are not clearly demarcated during encoding, as AC inference using the localist strategy requires clean conjunctive representations of direct AB and BC pairs (Kumaran & McClelland, 2012; Schapiro et al. 2017). In temporal statistical learning paradigms, objects are presented one at a time in a continuous stream with embedded regularities (Saffran et al., 1996). A localist strategy that quickly forms robust conjunctive representations of every observed temporal co-occurrence would encode both reliable and unreliable pairings. AC inference using this strategy would be very difficult, as recurrent processing at retrieval would activate the unreliable associations. In contrast, distributed representations are very well suited to capturing graded statistical regularities (Rogers & McClelland 2004; Schapiro et al. 2017). We therefore predicted that a statistical learning variant of our paradigm would show a robust advantage for interleaved exposure even in the slow explicit inference test.

In Exp 3a, we exposed participants to a sequence of objects in which object pairings were not clearly demarcated but instead defined by the relative frequency of consecutive occurrence (Fig. 4a). For each participant, a sequence of object pairs was generated following the same protocol as in Exps 1 and 2. During learning, however, objects appeared one at a time in a continuous stream, such that two objects from the same pair were always shown consecutively, followed by objects from a different pair. For each occurrence of an object pair, the order of the objects was randomized. As a cover task, participants were instructed to quickly judge whether the current object appeared heavier or lighter than the preceding object. After learning, participants were informed that there were object pairs embedded within the sequence they had viewed. Participants then completed the speeded recognition task followed by the explicit inference task as administered in Exps 1a-c. To ensure that participants were able to extract basic AB and BC associations during learning, we performed a replication study 3b with an additional explicit direct pair memory test.

As in Exps 1a-c, we observed evidence of stronger rapid recognition of interleaved relative to blocked indirect associations. Participants were slower at correctly rejecting interleaved relative to blocked ACs in 3a ($t(37)=2.54$, $p=0.016$; Fig. S7b) but not in 3b ($t(25)=-0.22$, $p=0.83$), whereas there was no difference between interleaved and blocked foils in 3a ($t(36)=-0.75$, $p=0.46$; Fig. S7b) or 3b ($t(27)=-0.16$, $p=0.87$). Participants exhibited higher false alarm rates for interleaved than blocked ACs in both 3a ($t(37)=5.82$, $p<0.001$; Fig. S7a) and 3b ($t(25)=2.31$, $p=0.03$). We found that interleaving induced a false sense of familiarity even for unrelated objects in this setting due to temporal proximity: response time was not significantly different between interleaved ACs and foils in 3a ($t(35)=1.91$, 0.064) or 3b ($t(26)=-0.17$, $p=0.87$), and participants showed higher false alarm rates for interleaved than for blocked foils in both 3a ($t(36)=4.07$, $p<0.001$) and 3b ($t(27)=2.70$, $p=0.012$).

Critically, in Exp 3a, unlike in Exp 1a-c, we observed higher accuracy for interleaved than blocked AC pairs in the explicit inference task ($t(42)=3.06$, $p=0.004$; Fig. 4b), despite no difference in direct pair accuracy ($t(42)=0.81$, $p=0.42$; Fig. S3a) or RT ($t(42)=-1.27$, $p=0.21$; Fig. S13) between conditions. We found significant above chance inference performance in the interleaved condition ($t(42)=3.25$, $p=0.002$) but chance performance in the blocked condition ($t(42)=-1.11$, $p=0.27$), suggesting that only interleaved exposure permitted successful inference (Fig. 4c). Exp 3b largely replicated these effects: performance in the explicit inference task was superior in the interleaved condition ($t(34)=2.09$, $p=0.044$; Fig. 4b) despite matched direct pair performance (speeded recognition accuracy: $t(34)=0.89$, $p=0.38$, Fig. S3a; speeded recognition RT: $t(34)=0.81$, $p=0.42$, Fig. S14; slow recognition accuracy: $t(34)=-0.95$, $p=0.35$, Fig. S3b; slow recognition RT: $t(33)=-0.45$, $p=0.66$, Fig. S14), indicating that the advantage in explicit inference for interleaved ACs was not due to a failure to learn direct AB and BC associations in the blocked condition (though note enhanced memory for blocked BC pairs in this case: $t(34)=-2.20$, $p=0.03$; Fig. S15). Consistent with the accuracy differences, participants in 3b responded faster to interleaved than to blocked ACs during explicit inference ($t(32)=-2.07$, $p=0.046$; Fig. S14; though this was not reliable in 3a: $t(42)=-1.12$, $p=0.27$; Fig. S13). As in Exp 1c, participants were slower during explicit recognition of indirect than direct pairs for both interleaved ($t(33)=4.77$, $p<0.001$; Fig. S9) and blocked trials ($t(32)=4.78$, $p<0.001$; Fig. S9). Together, the results indicate that in a setting where object pairings need to be inferred from the statistics of co-occurrence over time, interleaving benefits even explicit inference.

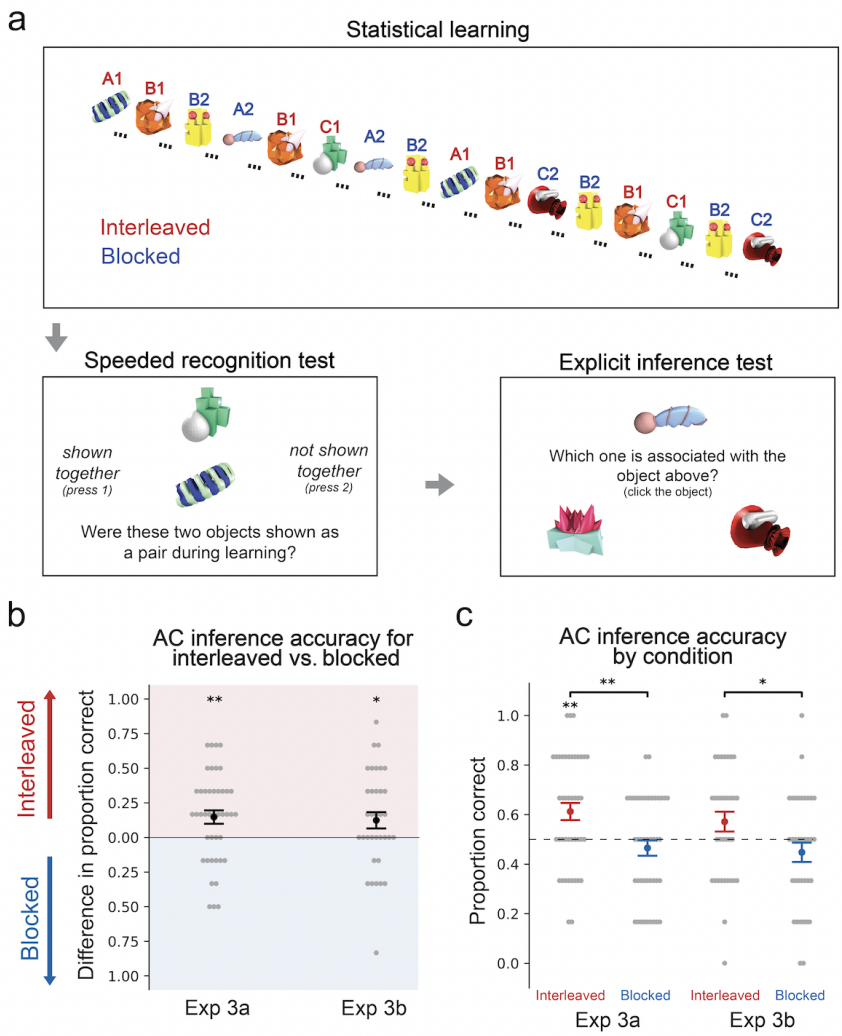

Figure 4. Exps 3a-b design and results. (a) Participants viewed objects presented one at a time followed by a speeded recognition test and explicit inference test. (b) Explicit inference accuracy differences between interleaved and blocked trials in Exp 3a and 3b. (c) Explicit inference accuracy by condition in Exp 3a and 3b.

Comparison of three models of associative inference. To evaluate how different classes of models behave in our paradigm, we contrasted three models of memory that have been proposed to solve associative inference (Fig. 5): the temporal context model (TCM; Howard & Kahana 2002; Howard et al., 2009; Fig. 5a-c), which employs distributed representations, REMERGE (Kumaran & McClelland 2012; Fig. 5d-f), which uses only localist representations, and our model of the hippocampus, C-HORSE (based on Schapiro et al., 2017; Fig. 5g-i), which contains both kinds of representation separated across the two pathways of the hippocampus. Code for these models can be found at: https://github.com/schapirolab/itlblklearning.

TCM encodes items by associating them with the contexts in which they are encountered. Relationships between items are then reflected in the degree of overlap in distributed context representations. TCM learns by updating item-to-context and context-to-item mappings. Prior to the experiment, distinct items are assumed to be associated with non-overlapping contexts. During learning, each input item causes drift in a context vector, and subsequent items are associated with this drifting context. Through this process, items' associated contexts become overlapping to the extent that items continue to appear in similar contexts.

To simulate the task, we adopted a version of TCM designed to simulate paired-associate learning (Howard et al., 2009; Supplementary Table 2). To delimit trials of pair presentations, a distractor vector drifts the context prior to each new trial. The context drift caused by an item is determined by the item representation and a recency-weighted average of past contexts of its occurrence. In the interleaved condition of our design, the context drift caused by each B will consistently reflect recent contexts in which its related A and C were presented. In contrast, the drift caused by each blocked item B will reflect only the contexts of its related A initially, until it becomes gradually swamped by that of its related C as BC pairs begin to appear. Prior context exponentially decays as new contextual information is incorporated. As a result of this interference, contexts in which interleaved A and C appear will be more similar than those of blocked A and C. Thus, in TCM, we predicted that interleaved ACs would develop stronger associations than blocked ACs.

In contrast to the distributed representations used in TCM, REMERGE (Kumaran & McClelland 2012; Fig. 5d; Supplementary Table 1) employs a localist code: AB and BC are represented as non-overlapping units in the conjunctive layer, each of which is bidirectionally connected to individual item units in the feature layer (i.e., AB in the conjunctive layer is connected to A and B in the feature layer). This model employs an idealized learning process whereby each pair presentation strengthens connections only to constituent items, such that each AB presentation only updates bidirectional A-AB and B-AB connections. Since weight updates are fully orthogonalized for different conjunctive pairs, there is no basis for a difference between interleaving and blocking in the model's behavior.

To infer item associations at retrieval, REMERGE engages in a search process by allowing activity to spread amongst connected units (McClelland et al., 1986). At the onset of retrieval, an external input activates the corresponding unit in the feature layer. At each timestep during retrieval, the net input into each unit is a weighted sum of its previous net input and the current inputs from connected units. In the feature layer, each unit's activity is a logistic function of its net input. In the conjunctive layer, a version of the softmax function normalizes net inputs across units, mimicking competitive inhibition. AC inference can be achieved when an input A causes activity to spread to AB, B, BC, and finally C. REMERGE is intended to simulate "big-loop" recurrence within the hippocampal

system, in which output activity from the hippocampus can be recirculated back into the system as input via the entorhinal cortex.

C-HORSE (based on Schapiro et al., 2017; Fig. 5g), inspired by known circuitry and properties of the area, posits that both distributed and localist representations are available via separate pathways. The model consists of an input layer representing superficial layers of entorhinal cortex (EC_in), an output layer representing deep layers of entorhinal cortex (EC_out), and hidden layers representing hippocampal subfields dentate gyrus (DG), cornu ammonis 3 (CA3), and CA1. We modified the architecture of our previously published model (Schapiro et al., 2017; see Supplementary Tables 3-5 for comparison) to incorporate anatomical evidence of topographic organization of projections from CA3 and superficial EC to CA1, along its proximodistal axis (Sun et al., 2014; Witter et al., 2006). We split the existing CA1 layer into two layers (proximal/distal CA1), such that layers in the model are linked via two different pathways: the trisynaptic pathway (TSP), EC_in -> DG -> CA3 -> pCA1 -> EC_out, and the monosynaptic pathway (MSP), EC <-> dCA1 (Fig. 5g). The model instantiates sparse connectivity and high inhibition in the TSP, which give rise to the well-known pattern-separated representations in DG and CA3 (Leutgeb et al., 2007), and dense connectivity and lower inhibition in the MSP, which produce relatively overlapping representations in dCA1. The TSP is also implemented with a faster learning rate than the MSP, consistent with evidence suggesting that the TSP can do one-shot learning, whereas the MSP does more incremental learning (Nakashiba et al., 2008). We expected that overlapping representations in the MSP would require interleaving to support inference, whereas pattern-separated representations in the TSP would form via either interleaved or blocked exposure and support inference via recurrence, as in REMERGE.

For TCM and the C-HORSE, we presented sequences of object pairs, in which AB and BC were either interleaved or blocked, generated using the same protocol and number of exposures as in Exps 1a-c. For REMERGE, we assumed that all connection weights are identical by the end of learning, as the extent of its weight strengthening depends only on the number of presentations, which is equivalent across all pairs. After learning, we measured the amount of activation of item A given C, and vice versa, across 100 instantiations of each model.

In TCM, as predicted, interleaving led to stronger AC activation relative to blocking ($t(99)=16.58$, $p<0.001$) and relative to matched foil pairs ($t(99)=13.29$, $p<0.001$) across learning (Fig. 5b) and at test (Fig. 5c). In addition, there was greater activation for interleaved foils than blocked foils (at the end of training: $t(99)=9.38$, $p<0.001$; Fig. 5b), reflecting the fact that contexts associated with adjacent trials are not fully orthogonalized. TCM can thus account for situations where interleaving benefits inference, but cannot explain matched performance, as observed in the explicit inference task in Exp 1.

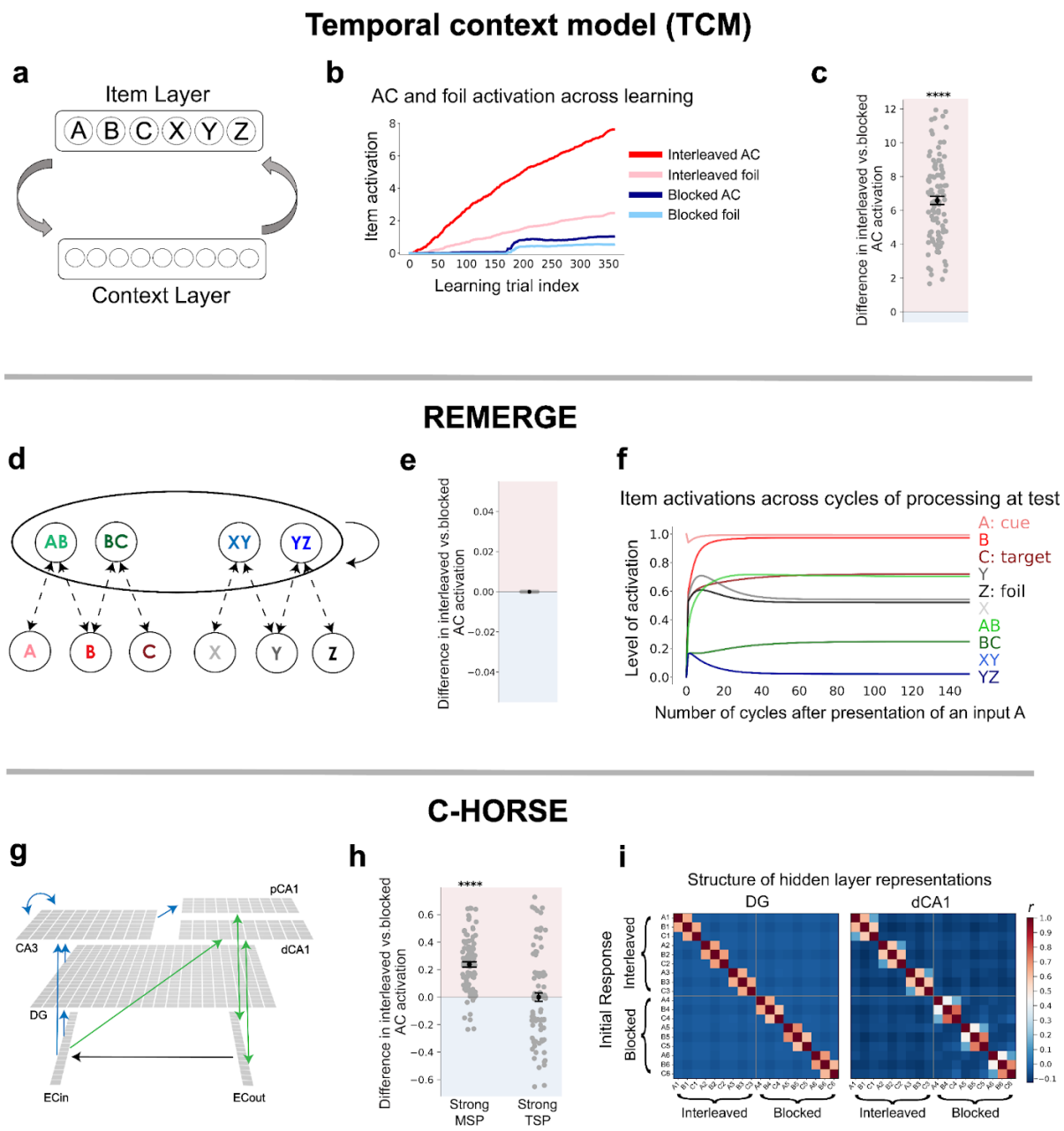

Figure 5. Model architectures and results. (a) TCM learns a distributed representation of the context associated with each item. **(b)** Throughout learning, indirectly related items A and C in the interleaved condition activate one another more than unrelated foils and those in the blocked condition. **(c)** Stronger activation of interleaved than blocked ACs at test across 100 simulations. **(d)** In REMERGE, AB and BC pairs are represented as idealized localist, conjunctive units connected to associated input units via bidirectional connections (dashed lines). **(e)** REMERGE predicts no difference between interleaved and blocked AC activation at test. **(f)** Item activations across a test trial after presentation of item A. As recurrent dynamics settle in the system, the activity of the target C is higher than that of matched foil Z. X is identical

to Z and XY is identical to YZ. **(g)** C-HORSE instantiates known connectivity between subfields of the hippocampus. **(h)** The model predicts an advantage for interleaved AC activation when behavior is driven predominantly by the MSP and equivalent AC activation between conditions when the TSP drives behavior. **(i)** Heatmap of hidden layer unit activity correlations in DG and dCA1 when individual items are presented. The initial response is used for these correlations, before activity spreads from EC_out to EC_in on a particular trial.

In REMERGE, given opportunities for spreading activation amongst related units, an input A always activates its associated items more than matched foils (Fig. 5f). This recurrent process first activates A's directly related B and AB units more than foils Y and XY, followed by greater activation of BC than of YZ. Finally, C becomes activated more than its matched foil Z. The model predicts no difference between interleaved and blocked AC activations (Fig. 5e), as learning in REMERGE depends on the number but not the order of presentations due to fully orthogonalized weight updates across conjunctive pairs. REMERGE thus provides an account of how inference performance can be matched between conditions, but fails to explain the observed advantage of interleaved associations in the implicit, speeded task.

In C-HORSE, we simulated the speeded and explicit inference tasks by modulating reliance on the MSP and TSP. We assumed a control mechanism that shifts the relative strength of dCA1/pCA1 outputs to EC_out during AC inference such that model dynamics were predominantly driven by only one of the two pathways at a time. Such a mechanism could potentially be implemented in interactions between medial prefrontal cortex and CA1, as medial prefrontal cortex is known to influence CA1 representations as a function of task requirements (Guise & Shapiro, 2017; Eichenbaum, 2017). We found that when relying primarily on the sparse, pattern-separated representations of the TSP (Fig. 5i), the model demonstrates matched performance between the interleaved and blocked conditions ($t(99)=-0.03$, $p=0.98$; Fig. 5h), akin to human participants on the explicit inference task. On the other hand, relying on the overlapping representations of the MSP (see overlap between A and C items in Fig. 5i, dCA1) leads to an advantage for the interleaved condition ($t(99)=12.17$, $p<0.001$; Fig. 5h), as in human performance on the speeded recognition task.

In sum, similar to generic neural network models employing distributed representations (McClelland et al., 1995), TCM predicts an advantage of interleaved exposure for linking ACs. TCM shares with neural network models the use of distributed representations of item associations (mediated by shared context), rendering it similarly sensitive to interference. Unlike generic neural network models and TCM, REMERGE employs an idealized localist representation that is insensitive to the ordering of input presentations and equally supports interleaved and blocked AC inference. C-HORSE employs both kinds of representations: representations in the MSP are overlapping as those in

generic neural network models and in TCM, whereas the TSP hosts pattern-separated representations similar to those in REMERGE. We propose that both representations exist in the hippocampus, and that both are needed in order to account for the pattern of behavior we observed empirically.

Discussion

A hallmark of successful neural network models is the processing of related information using overlapping populations of neuronal units. These distributed representations have been posited to be employed throughout the brain, especially in the neocortex (McClelland et al., 1995). In neural network models, learning in interleaved order is fundamental to building distributed representations that support useful behavior. Does interleaved exposure similarly facilitate the formation of distributed representations in humans? We previously put forward a model proposing that the CA1 subfield of the hippocampus hosts distributed representations learnable on the timescale of one experimental session (Schapiro et al., 2017). We thus tested this idea using the hippocampally-dependent associative inference task (Bunsey & Eichenbaum, 1996; Preston et al., 2004), assessing participants' ability to link indirect associations (AC) after interleaved or blocked exposure to directly associated items (AB and BC). We found that participants' behavior reflected the use of distributed representations: after interleaved learning, participants exhibit an increased capacity to rapidly recognize item relatedness, to efficiently generalize novel information based on learned associations, and to make inferences according to statistical regularities of item associations.

Models employing distributed representations support efficient judgments of item relatedness via directly overlapping representations, whereas localist models require spreading activation processes to support such judgments (McClelland et al., 1986). We postulated that if distributed representations are available for learning on a short timescale, interleaving AB and BC pairs across an experimental session would drive the formation of overlapping representations of A and C. These representations would result in higher sensitivity to the AC association in a speeded, implicit setting where spreading activation processes might be constrained. In a scenario with opportunity for additional explicit retrieval-based processing over pattern-separated codes, we would expect no difference between conditions. Indeed, a previous study identified no behavioral difference between interleaving and blocking when participants were asked to deliberately infer associations between indirectly related items (Schlichting et al., 2015). As in this prior study and as predicted, in Exp 1, there was no difference between blocked and interleaved conditions in the standard explicit inference task. In an implicit, speeded task, however, there was significant slowing in the response to interleaved ACs—it was more difficult to indicate that these items had not been studied together.

Participants also tended to false alarm to the interleaved ACs, believing that they had studied these items together directly. These results suggest that interleaving supports direct AC recognition, whereas blocked associations require additional retrieval-based processes. The findings are consistent with the idea that associative retrieval-based processes are slower and more explicit than those underlying recognition (Nobel & Shiffrin 2001; Cohn & Moscovitch 2007).

In the context of a recognition task, interleaving *impaired* performance, leading participants to indicate that they recognized pairs of items that were never studied together. But we expect that interleaving should often benefit behavior. If interleaving builds distributed representations, it should afford automatic generalization of new knowledge amongst related entities. In Exp 2, after learning AB and BC associations, participants learned novel arbitrary attributes of some items, and we assessed their ability to generalize such attributes between interleaved and blocked ACs. Generalizing novel associations to indirectly related items was superior in the interleaved condition and was not above chance for blocked items, consistent with the idea that interleaving builds up distributed representations that support subsequent generalization.

Across the first two experiments, interleaved associations benefitted performance in tasks that required relatively rapid, implicit judgments. Are there situations where interleaving would benefit performance even under more explicit conditions? We hypothesized that a statistical learning paradigm (Saffran et al., 1996) would provide such a situation, where direct associations need to be learned over time from graded co-occurrence frequencies. Distributed representations are especially sensitive to this kind of graded statistical information—indeed, our model indicates a complete failure of localist representations in the hippocampus to support this kind of learning (Schapiro et al. 2017). In Exp 3, we exposed participants to objects with the same pair structure, but presented one at a time in a continuous sequence, with each object temporally adjacent to its pairmate. Consistent with our prediction of a qualitative advantage of interleaved associations in this scenario, performance was superior in the interleaved condition during the explicit inference task, with inference no different than chance in the blocked condition. The results suggest that interleaving is essential for forming representations that permit inference in a scenario that requires integrating statistical information across time.

Simulations contrasting different computational models of human memory provided evidence that at least two forms of representation are likely to be at play in human inference, with interleaving especially beneficial for the formation of distributed representations. We found that using exclusively distributed or exclusively localist internal item representations rendered a model unable to account for the data: TCM (Howard & Kahana 2002; Howard et al., 2009), employing a distributed code, predicts an advantage of interleaving over blocking but fails to account for the possibility of equivalent performance (as in explicit inference). REMERGE (Kumaran & McClelland 2012), a model that is exclusively localist, accounts for the matched performance, but never predicts the

observed interleaving advantage. Our model of the hippocampus, employing both kinds of representations via different pathways, can account for both phenomena.

There are many domains where interleaving has proven broadly beneficial for learning (Brunmair & Richter 2019), such as in educational settings (Taylor & Rohrer 2010) and in motor skill learning (Goode & Magill 1986). There is also extensive prior research on the effects of interleaved and blocked exposure on learning categories of multidimensional stimuli (see Carvalho & Goldstone, 2015 for a review). Though interleaving stimuli from different categories more often benefits learning and generalization (Birnbaum et al., 2013; Kang & Pashler, 2012; Kornell & Bjork, 2008; Noh, Bjork, & Preston, 2021), this is modulated by several factors, such as the kind of stimuli, the structure of the categories, and task instructions. Many of these findings have been attributed to trial-by-trial attentional effects to different features: Exposure to stimuli from the same category back-to-back (blocking) promotes attention to within-category similarities, whereas exposure to stimuli from different categories back-to-back (interleaving) promotes attention to across-category differences (Carvalho & Goldstone, 2015; Carvalho & Goldstone, 2017). Our study was designed to avoid these kinds of attentional effects: Attention to specific features in adjacent trials neither benefited nor harmed learning, as there were no shared features across triads. The benefit for interleaving in our data is thus more likely to be related to the benefit observed in neural network models, which do not (typically) have these trial-by-trial attentional biases.

One recent category learning paper found better performance under blocked than interleaved conditions, which the authors argued was supported by 'factorized' representations that they suggest may be difficult to explain under an attentional account (Flesch et al., 2018). The authors interpret the results as evidence that neocortex may not be as susceptible to catastrophic interference as neural network models would predict. An alternative interpretation, consistent with the current framework, is that orthogonalized representations in areas DG and CA3 of the hippocampus learn factorized representations and thus can support behavior in scenarios where it is not advantageous to integrate across conditions, as was the case in that study.

In addition to the distributed and localist strategies considered above, another influential proposal for how the hippocampus may carry out associative inference is known as 'integrative encoding' (Shohamy & Wagner, 2008; Schlichting & Preston 2015). Integrative encoding posits that studying BC, after having studied AB, triggers reinstatement of the AB memory through pattern completion mechanisms, and an overlapping representation of AB and BC is then encoded. Although this strategy, similar to our account, employs overlapping AC representations formed during encoding to support inference, it relies on the episodic encoding and pattern completion mechanisms of DG and CA3. It is not clear whether integrative encoding predicts an advantage for blocked or interleaved presentation: Blocking may lead to strong AB memory that permits better reinstatement during BC learning, or temporal proximity between related AB and BC presentations during interleaving may

promote the formation of an overlapping AB-BC representation (Schlichting et al., 2015). Future work using computational models that implement integrative encoding could fruitfully explore these possibilities and their relationship to our findings.

Several prior imaging and behavioral studies speak to the different strategies and representations that can underlie associative inference. Some have argued that the hippocampus constructs overlapping representations of related AB and BC pairs at *encoding* to support later inference, as in integrative encoding (Shohamy & Wagner 2008; Zeithamova et al., 2012; Schlichting et al., 2014; Schlichting et al., 2015), whereas others have argued that inference is supported by sequential activation at *retrieval* (Banino et al., 2016; Koster et al., 2018; Barron et al., 2020; de Araujo Sanchez & Zeithamova 2020). This dichotomy mirrors the distinction in decision-making between model-free and model-based behavior (Daw et al., 2005), where information is either directly stored in representations during learning or computed on the fly. We demonstrate that repeated, interleaved exposure facilitates the encoding of overlapping representations. This is in line with existing evidence that encoding-based mechanisms support inference after interleaved exposure (Shohamy & Wagner 2008; Zeithamova et al., 2012), whereas inference may rely on sequential, retrieval-based processes when related pairs are shown in a blocked manner (Barron et al., 2020; Banino et al., 2016) or with limited exposure (Banino et al., 2016; Koster et al., 2018). In both our explicit and implicit assessments, we found faster RTs for direct than indirect pairs, which has sometimes been interpreted as evidence for retrieval-based mechanisms (Shohamy & Wagner 2008; de Araujo Sanchez & Zeithamova 2020). However, models learning distributed representations at encoding also predict stronger associations among directly learned associates, so we do not view these RT differences as diagnostic between the mechanisms considered here. In potential tension with our results, an fMRI study (Schlichting et al., 2015) found that blocked exposure promoted apparent integration of related pairs in the anterior hippocampus. However, encoding- and retrieval-based inference strategies may produce similar results in the slow fMRI BOLD signal: through spreading activation, retrieval-based strategies will activate overlapping sets of items for related inputs even if their underlying representations are not integrated (Kumaran et al., 2012; Chen et al., 2021).

Taken together, we provide converging behavioral and computational evidence that interleaved exposure facilitates the formation of distributed representations in humans. The rapid timescale of their formation in this paradigm—across one experimental session—demonstrates the potential for tractable further investigation of this powerful form of representation.

Methods

Experiments 1a-c

Participants. In Experiment 1a, we recruited 33 participants (9 females, 1 unknown; 24 White, 2 Black, 1 American Indian/Native, 4 Asian, and 2 unknown; 2 Hispanic or Latino; mean age = 34.75, SD = 9.57), with 26 participants remaining after exclusions. Experiment 1b was a pre-registered study (https://osf.io/ag42z) for which 83 participants were recruited (35 females, 3 unknown; 57 White, 9 Black, 2 American Indian/Native, 11 Asian, and 4 unknown; 6 Hispanic or Latino; mean age = 37.72, SD = 10.42), with 54 participants after exclusions. Experiment 1c was an additional replication study in which we recruited a total of 184 participants (90 females, 8 unknown; 144 White, 16 Black, 3 American Indian/Native, 13 Asian, and 8 unknown; 13 Hispanic or Latino, 3 unknown; mean age = 36.64, SD = 11.56), with 100 participants after exclusions. We collected additional participants in this study for the purposes of analyses unrelated to this paper. Across all three experiments, we excluded participants with d-prime lower than 1.5 on the speeded recognition task, indicating poor memory for the studied direct pairs. This exclusion criterion departed from our pre-registered version, which only considered the hit rate in the speeded task, as we realized that it is also important to identify participants with high false alarm rates (participants who failed to reject objects that were neither directly nor indirectly related). Participants across all experiments were recruited through Amazon Mechanical Turk. The study protocol was approved by the local Institutional Review Board.

Design and procedure. We adopted a within-subject design in which, during learning, each participant was shown a sequence consisting of presentations of 12 pairs of novel visual objects randomly sampled from 36 artificial object images (Hsu et al., 2014; Schlichting et al., 2015). Each "direct" pair AB was uniquely related to another direct pair BC through a linking item B. Each direct pair was shown 30 times. Among the six ABC triads shown to each participant, three were interleaved, with AB and BC appearing in alternation, and three were blocked, with all presentations of AB occurring before the first presentation of BC. Presentation of the two pair types was intermixed throughout the learning phase, and there was no demarcation of the halfway point. Pairs that share an object were never shown back-to-back. Participants were instructed to remember the pairings of objects by creating quick narratives of how two objects might interact. Each participant completed a total of 720 trials, with the two objects displayed side-by-side on the screen for 1000 ms. Across repeated presentations of an object pair AB, the horizontal position of A relative to B was randomized. Subsequent to each pair presentation, participants saw the question "on a scale of 1 (failed to visualize a story) to 5 (clearly visualized a story), how well were you able to visualize a story linking the objects?" Participants responded with a numerical rating with a 7000 ms response deadline in Exps 1a and 1b, and a 2000 ms response deadline in 1c.

After learning, participants completed two tasks that probed their memory of learned object associations: a speeded recognition task followed by an explicit inference task (Fig. 1). During the speeded recognition task, on each trial, two objects were displayed for 1500 ms and participants were asked to respond within 3500 ms whether the two objects had been shown as a pair during the learning phase. The task comprised 24 trials with directly paired objects (AB), 12 trials with indirectly related objects (AC), and 12 trials with unrelated foil pairs (e.g., AD). Each object pair appeared in two trials, with two different vertical positions (A above C or C above A), with the constraint that pairs could not be repeated on back-to-back trials. At the onset of the explicit inference task, participants were instructed that two items paired with the same item form an indirect association (i.e., A and C items paired with the same B) and that they had to identify indirectly related items in this task. On each trial, participants saw a cue object at the top of the screen (A or C) and were instructed to select which of two objects shown below the cue object was indirectly related to it. The foil was of the same type as the target (and from the same interleaved or blocked condition) but from a different triad. There was a 7000 ms response deadline. The explicit inference task consisted of 12 trials where each AC association was tested twice, such that each item served as the cue object in one trial and as the target object in the other.

In Exp 1c, participants then completed an additional task that employed the same design as the explicit inference task, assessing memory for direct AB and BC pairs. This task consisted of 24 trials: 12 AB and BC pairs each shown twice, with each object shown as the cue in one trial and the target in the other. On each trial, participants selected between a target object and a matched foil with a 7000 ms response deadline.

Analysis. For each RT analysis, we excluded participants who had no data recorded for a condition of interest (e.g., missed all interleaved AC trials for the analysis of interleaved – blocked RT difference). Across all experiments, RTs for trials during which participants responded correctly were log-transformed before averaging. For each variable of interest, we performed paired t-tests to test the significance of within-subject differences between two conditions of interest.

Experiment 2

Participants. In Experiment 2, 84 participants (33 females, 3 unknown; 67 White, 9 Black, 0 American Indian/Native, 2 Asian, and 6 unknown; 14 Hispanic or Latino; mean age = 39.26, SD = 11.55) were recruited. We excluded participants who missed more than half of explicit direct pair memory trials or more than one third of property memory trials, resulting in a total of 53 participants.

Design and procedure. After learning direct pair associations as in Exps 1a-c, participants learned novel properties of a subset of the objects, and then completed a property memory and property

generalization task. During property learning, each participant either learned to associate the six A objects (25 participants) or the six C (26 participants) objects across both conditions with six novel nonwords, such as "flises," selected from the ARC nonword database (Rastle et al., 2002). Prior to property learning, we instructed participants that each of the objects they saw had a hidden property and that they would learn some of these properties. On each property learning trial, an object was displayed for 2500 ms along with a text description of its associated property ("this object flies"). After each object-nonword association presentation, participants rated on a scale of 1 to 5 how well they felt they were able to learn the association, with a 4000 ms response deadline. Each object-property association was displayed 12 times during property learning.

During the property memory task, on each trial, participants saw a nonword property at the top of the screen (e.g., "flises") and were instructed to select, between a target object and a matched foil, the object that was associated with the nonword during property learning. There was a 4000 ms response deadline. Each property association was tested twice, with two different horizontal positions (i.e., target on the left and foil on the right, and vice versa), for a total of 12 trials.

Before the property generalization task, participants were instructed that they would see a property they had studied along with two objects that they had not learned hidden properties for, and they would be asked to judge which of the two objects has that property. On each trial, participants saw a property at the top of the screen (e.g., "flises") and selected between a target object and a matched foil, with a 4000 ms response deadline. The target object was always indirectly related to the associated object (i.e. A to C). Each object-nonword association was tested twice, with two different horizontal positions (i.e., target on the left and foil on the right, and vice versa), forming a total of 12 trials.

Experiments 3a-3b

Participants. In Experiment 3a, 105 participants (51 female, 1 unknown; 85 White, 8 Black, 1 American Indian/Native, 3 Asian, and 8 unknown; 9 Hispanic or Latino; mean age = 37.74, SD = 11.58) were recruited through Amazon Mechanical Turk. We excluded participants who missed more than half of learning phase responses or more than one third of speeded direct trials, resulting in 43 participants. Experiment 3b was a replication study in which 100 participants (36 female, 3 unknown; 83 White, 12 Black, 1 American Indian/Native, 3 Asian, and 1 unknown; 9 Hispanic or Latino; mean age = 37.80, SD = 11.27) were recruited, resulting in 35 participants after exclusions.

Design and procedure. For each participant, a sequence of visual object pairs was generated following the same protocol as Exps 1a-c, except that each pair repeated 24 (instead of 30) times. During learning, however, each participant saw objects presented one at a time in a continuous stream with no breaks, such that two objects from the same pair were shown consecutively followed

by objects from a different pair. For each occurrence of an object pair AB (or BC), the order of the objects was randomized. Therefore, for an object pair AB or BC, the two objects always appeared adjacent in the sequence, but each object would also appear adjacent to objects that were neither directly nor indirectly related. As a cover task, participants were instructed to quickly respond as to whether the current object appeared heavier than the previous object in the sequence. Participants pressed one key if the current object seemed heavier than the preceding object, and a different key if not. Each object was displayed on the screen for 1000 ms followed by an inter-trial interval of 500 ms. After learning, participants were informed that there were object pairs embedded within the sequence they saw. Participants then completed the speeded recognition task followed by the explicit inference task as in Exps 1a-c. In 3b, participants additionally completed an explicit direct pair recognition task as in Exps 1c and 2.

Data Availability

The data analyzed in this study can be found at: https://github.com/schapirolab/itlblklearning.

References

Banino, A., Koster, R., Hassabis, D., & Kumaran, D. (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports*, *6*(1), 1–15.

Barker, G. R., Banks, P. J., Scott, H., Ralph, G. S., Mitrophanous, K. A., Wong, L.-F., Bashir, Z. I., Uney, J. B., & Warburton, E. C. (2017). Separate elements of episodic memory subserved by distinct hippocampal–prefrontal connections. *Nature Neuroscience*, *20*(2), 242–250.

Barron, H. C., Reeve, H. M., Koolschijn, R. S., Perestenko, P. V., Shpektor, A., Nili, H., Rothaermel, R., Campo-Urriza, N., O'Reilly, J. X., & Bannerman, D. M. (2020). Neuronal computation underlying inferential reasoning in humans and mice. *Cell*, *183*(1), 228–243.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392–402.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, *379*(6562), 255–257.

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 505.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699.

Chen, X., Martin, R., & Fischer-Baum, S. (2021). Challenges for using Representational Similarity Analysis to Infer Cognitive Processes: A Demonstration from Interactive Activation Models of Word Reading. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Cohn, M., & Moscovitch, M. (2007). Dissociating measures of associative memory: Evidence and theoretical implications. *Journal of Memory and Language*, *57*(3), 437–454.

de Araujo Sanchez, M. A., & Zeithamova, D. (2020). *Generalization and source memory in acquired equivalence*.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*(8), 2051–2062.

Doumas, L. A., & Hummel, J. E. (2005). Approaches to modeling human mental representations: What works, what doesn't and why. *The Cambridge Handbook of Thinking and Reasoning, Ed. KJ Holyoak & RG Morrison*, 73–94.

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322.

Goode, S., & Magill, R. A. (1986). Contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport*, *57*(4), 308–314.

Guise, K. G., & Shapiro, M. L. (2017). Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding. *Neuron*, *94*(1), 183–192.

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hinton, G. E. (1984). Distributed representations. *Tech. Rept. CMU-CS-84-157, Carnegie-Mellon University, Computer Science Department, Pittsburgh, PA*.

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: Transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 391.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Hsu, N. S., Schlichting, M. L., & Thompson-Schill, S. L. (2014). Feature diagnosticity affects representations of novel and familiar objects. *Journal of Cognitive Neuroscience*, *26*(12), 2735–2749.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592.

Koster, R., Chadwick, M. J., Chen, Y., Berron, D., Banino, A., Düzel, E., Hassabis, D., & Kumaran, D. (2018). Big-loop recurrence within the hippocampal system supports integration of information across episodes. *Neuron*, *99*(6), 1342–1354.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573.

Leutgeb, J. K., Leutgeb, S., Moser, M.-B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science*, *315*(5814), 961–966.

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M.-B., & Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, *305*(5688), 1295–1298.

McClelland, J. L., McNaughton, B. L., & Lampinen, A. K. (2020). Integration of new information in memory: New insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B*, *375*(1799), 20190637.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

McClelland, J. L., Rumelhart, D. E., & Group, P. R. (1986). *Parallel distributed processing* (Vol. 2). MIT press Cambridge, MA.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Nakashiba, T., Young, J. Z., McHugh, T. J., Buhl, D. L., & Tonegawa, S. (2008). Transgenic inhibition of synaptic transmission reveals role of CA3 output in hippocampal learning. *Science*, *319*(5867), 1260–1264.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 384.

Noh, S. M., Bjork, R. A., & Preston, A. (2021). *General knowledge and detailed memory benefit from different training sequences*.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, *51*(1), 78–109.

Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus*, *14*(2), 148–152.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362.

Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience*, *13*(1), 87–103.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*(1), 1–10.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8.

Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA1 subfield contributions to memory integration and inference. *Hippocampus*, *24*(10), 1248–1260.

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*(6846), 549–553.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389.

Sun, Y., Nguyen, A. Q., Nguyen, J. P., Le, L., Saur, D., Choi, J., Callaway, E. M., & Xu, X. (2014). Cell-type-specific circuit connectivity of hippocampal CA1 revealed through Cre-dependent rabies tracing. *Cell Reports*, *7*(1), 269–280.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848.

Witter, M. P., Wouterlood, F. G., Naber, P. A., & Van Haeften, T. (2000). Anatomical organization of the parahippocampal-hippocampal network. *Annals of the New York Academy of Sciences*, *911*(1), 1–24.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179.

Supplementary Information

Supplementary Table 1 - Parameters for simulations using REMERGE

parameter	*description*	*value*
w	the weight of bidirectional connections	1.52
τ	the temperature parameter for the logistic function and the hedged softmax function	0.4
C	the constant term for the hedged softmax function	1
λ	the proportion of the net input to a unit given by that on the previous timestep	0.2

Supplementary Table 2 - Parameters for simulations using TCM

parameter	*description*	*value*
γ_{fc}	the parameter that controls the learning rate for the feature-to-context matrix	0.1
γ_h	the proportion of t^{IN} given by h associated with the input item	0.9
β_{enc}	the rate of context drift during encoding produced by each t^{IN}	0.7
$\beta_{distractor}$	the rate of context drift produced by each distractor	0.99
γ_{cf}	the rate of update for the context-to-feature matrix	0.5

Supplementary Table 3 - Layer parameters for neural network model of the Hippocampus, as modified from Schapiro et al. (2017)

Layer	**Parameter**	**Schapiro et al. (2017)**[a]	**Current Model**[b]
pCA1	Layer	Does not exist	Added
	# Units	Does not exist	50
	Inhibition	Does not exist	FFFB (Gi = 3)
dCA1	Layer	Does not exist	Added
	# Units	Does not exist	50
	Inhibition	Does not exist	FFFB (Gi = 2.2)
CA1	Layer	Exists	Removed
	# Units	100	-
	Inhibition	kWTA Avg Inhib 0.25 pct/0.7 pt	-
DG	Inhibition	kWTA Avg Inhib 0.01 pct/0.9 pt	FFFB Gi = 24
CA3	Inhibition	kWTA Avg Inhib 0.06 pct/0.7 pt	FFFB Gi = 4.5
EC_in and EC_out	# Units	8/15/9	18
	Inhibition	kWTA Inhib k = 2/0.5 pt	FFFB Gi = 2.0

a The Schapiro et al. (2017) model was developed in C++ Emergent v.7.0.1 (7740) (http://grey.colorado.edu/emergent)

b The current model was developed in Golang Emergent v.1.0.5 (https://github.com/emer/emergent)

Supplementary Table 4 - Projection parameters for neural network model of the Hippocampus, as modified from Schapiro et al. (2017)

Projection	Parameter	Schapiro et al. (2017)	Current Model
Input -> EC_in	Weight range	0.25-0.75	0.8
EC_in -> DG	Learning rate	0.2	0.4
EC_in -> CA3	Learning rate	0.2	0.4
CA3 -> CA3	Learning rate	0.2	0.4
	WtScale.Rel	1	2
CA3 -> pCA1/CA1	Connectivity	100%	0.25
EC_in -> dCA1/CA1	WtScale.Abs	3	1
	Learning rate	0.02	0.04
dCA1/CA1 -> EC_out[a]	Learning rate	0.02	0.04
pCA1/CA1 -> EC_out[a]	Learning rate	0.02	0.04
EC_out -> dCA1/CA1	Learning rate	0.02	0.04
EC_out -> pCA1/CA1	Learning rate	0.02	0.04
EC_out -> EC_in	Weight range	0.49 - 0.51	0.7
	WtScale.Abs	2	1
	WtScale.Rel	0.5	1

[a] See Learning Table for changes in Absolute Weight Scaling.

Supplementary Table 5 - Learning Weight Scale parameters for neural network model of the Hippocampus, as modified from Schapiro et al. (2017)

Train									
Schapiro et al. (2017)[a]					**Current Model[b]**				
Projection	**Q1**	**Q2**	**Q3**	**Q4**	**Projection**	**Q1**	**Q2**	**Q3**	**Q4**
Ecin -> CA1	1	0	0	1	Ecin -> dCA1	1	0	0	2
CA3 -> CA1	0	1	1	0	CA3 -> pCA1	0	1	1	1
CA1 -> Ecout	4	4	4	4	dCA1 -> Ecout	1	0	0	1
					pCA1 -> Ecout	0	1	1	1

Test									
Schapiro et al. (2017)[a]					**Current Model[b]**				
Projection	**Q1**	**Q2**	**Q3**	**Q4**	**Projection**	**Q1**	**Q2**	**Q3**	**Q4**
Ecin -> CA1	1	0	0	1	Ecin -> dCA1	1	1	1	1
CA3 -> CA1	0	1	1	0	CA3 -> pCA1	1	1	1	1
CA1 -> Ecout	4	4	4	4	dCA1 -> Ecout[c]	(4)1	(4)1	(4)1	(4)1
					pCA1 -> Ecout[c]	1(4)	1(4)	1(4)	1(4)

[a] No. of cycles per quarter - Q1: 40, Q2: 20, Q3: 20, Q4: 20.

[b] No. of cycles per quarter - Q1: 25, Q2: 25, Q3: 25, Q4: 25.

[c] Symmetric upregulation of Absolute Weight Scale Values (1 vs 4)

Supplementary Figures

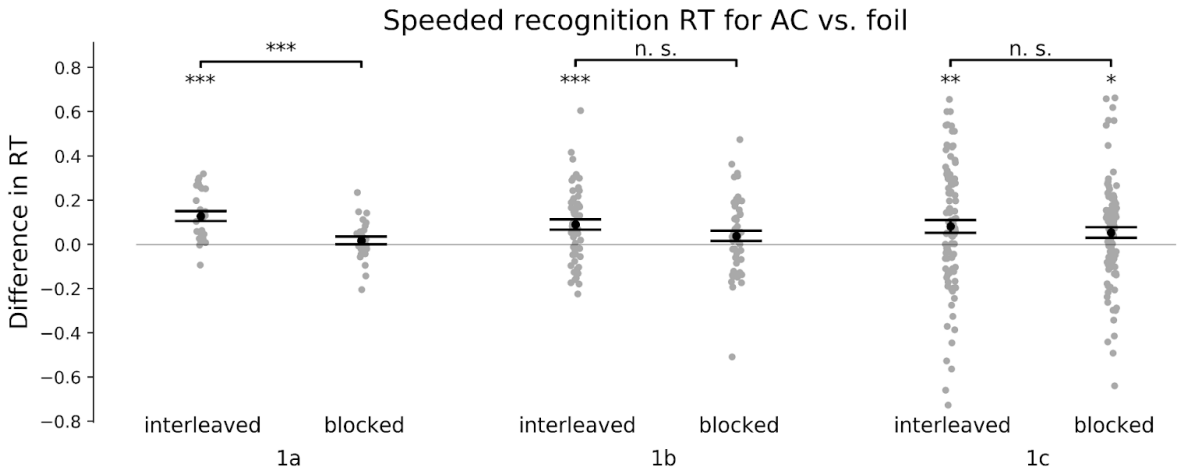

Figure S1. RTs for ACs — foils during speeded recognition in Exps 1a-c.

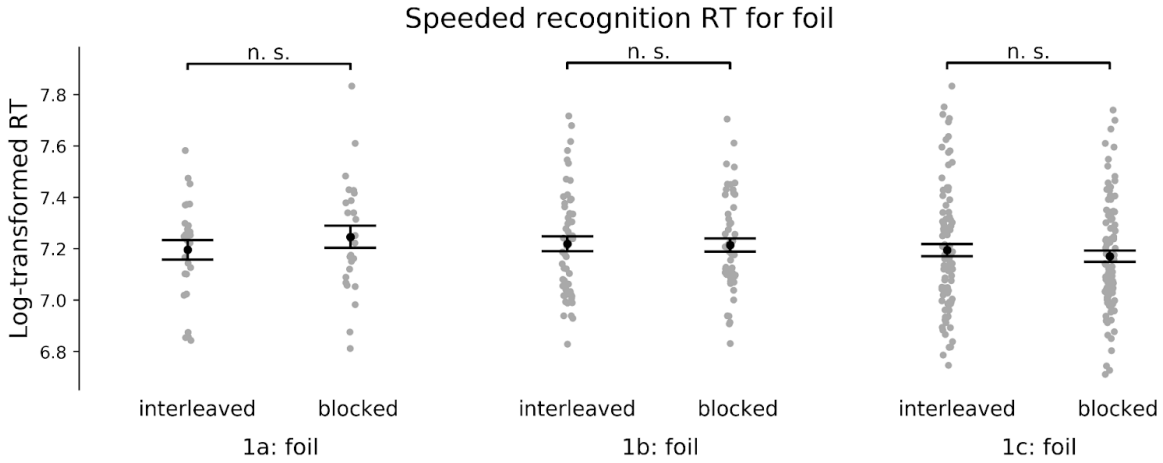

Figure S2. RTs for foil trials during speeded recognition in Exps 1a-c.

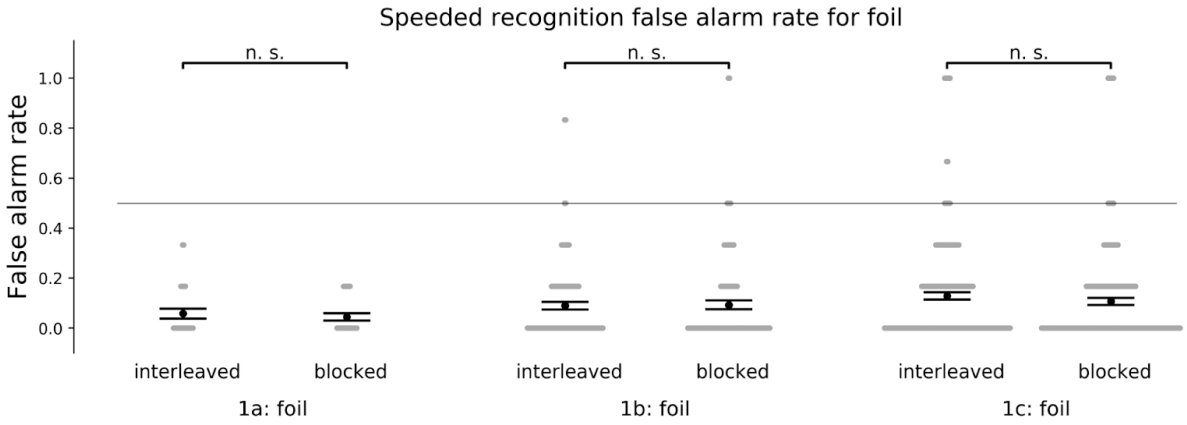

Figure S3. False alarm rates for foil trials during speeded recognition in Exps 1a-c.

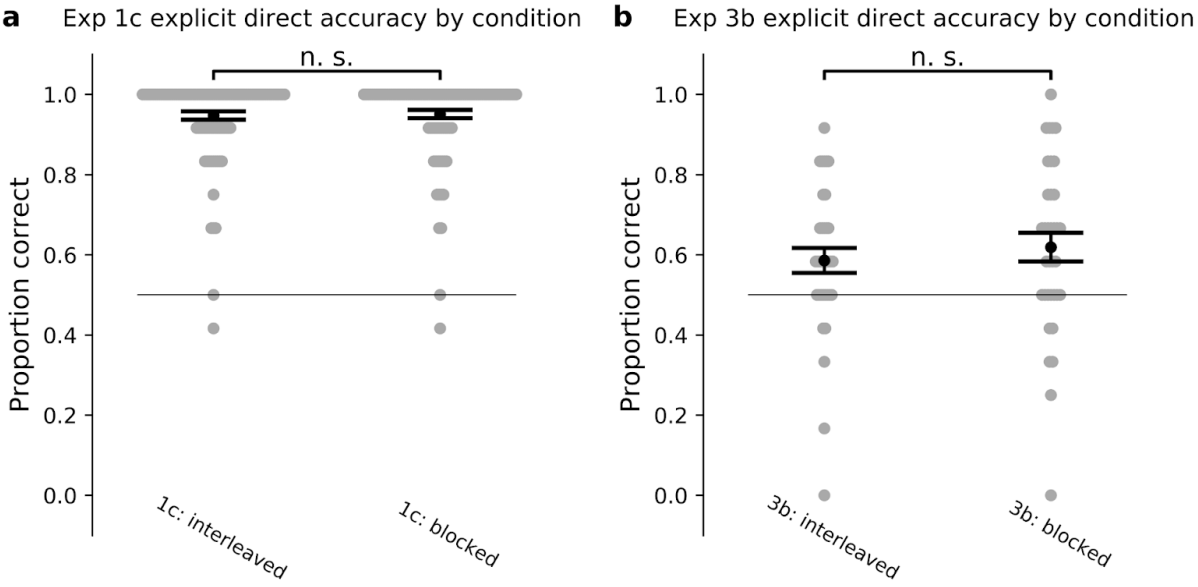

Figure S4. Accuracy for explicit direct trials in Exps 1c and 3b.

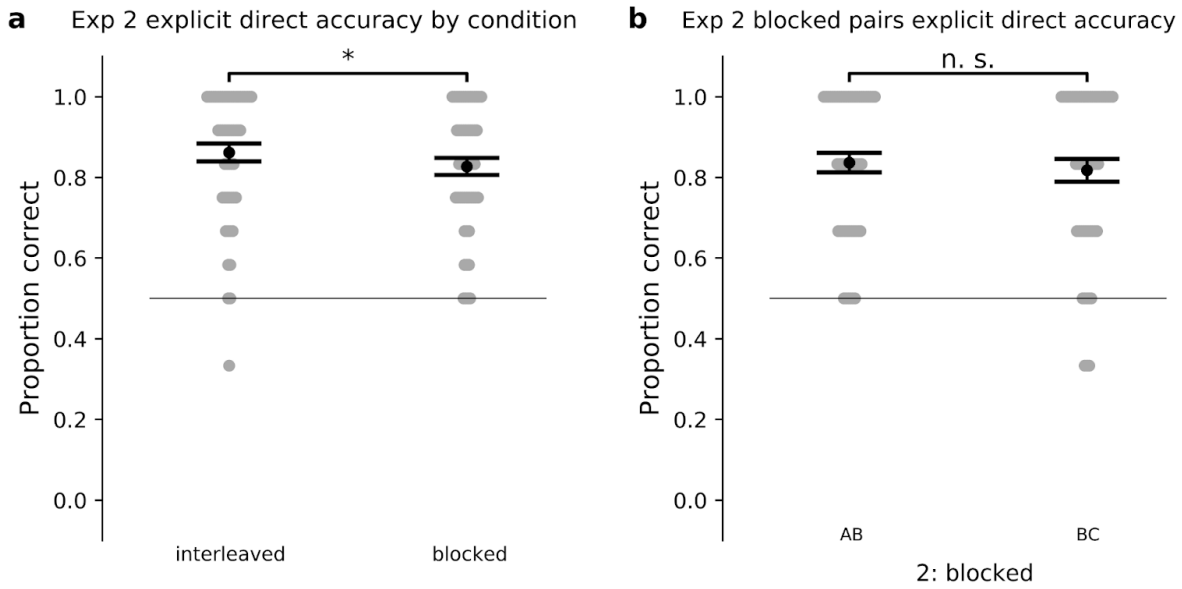

Figure S5. Accuracy for explicit direct trials in Exp 2.

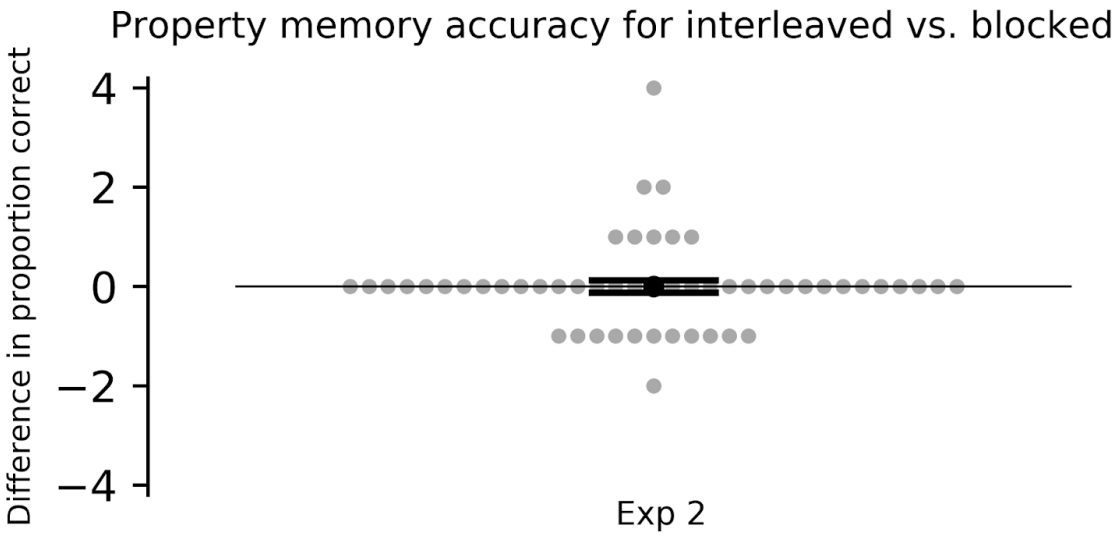

Figure S6. Property memory accuracy for interleaved vs. blocked trials in Exp 2.

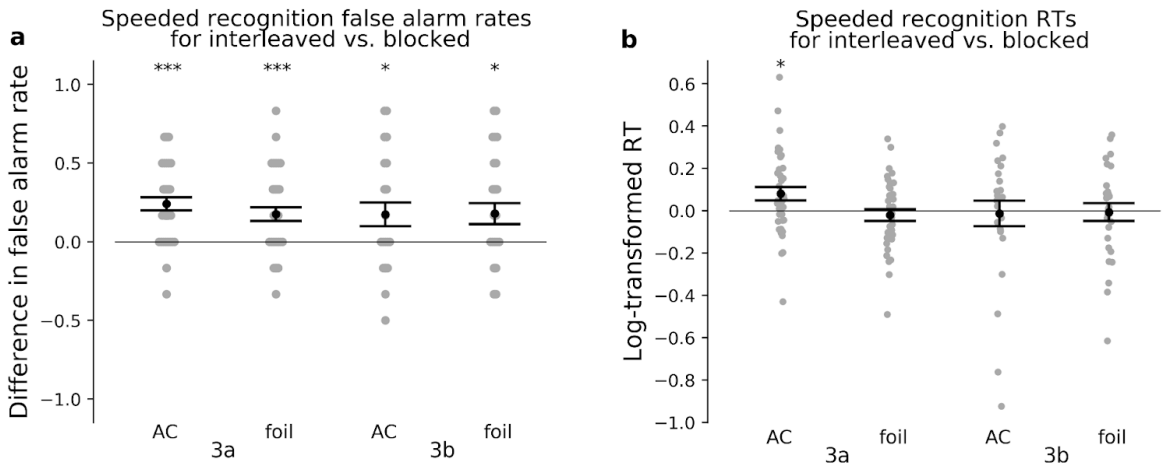

Figure S7. (a) False alarm rates for interleaved — blocked AC and foil trials during speeded recognition in Exps 3a-b. (b) RTs for interleaved — blocked AC and foil trials during speeded recognition in Exps 3a-b.

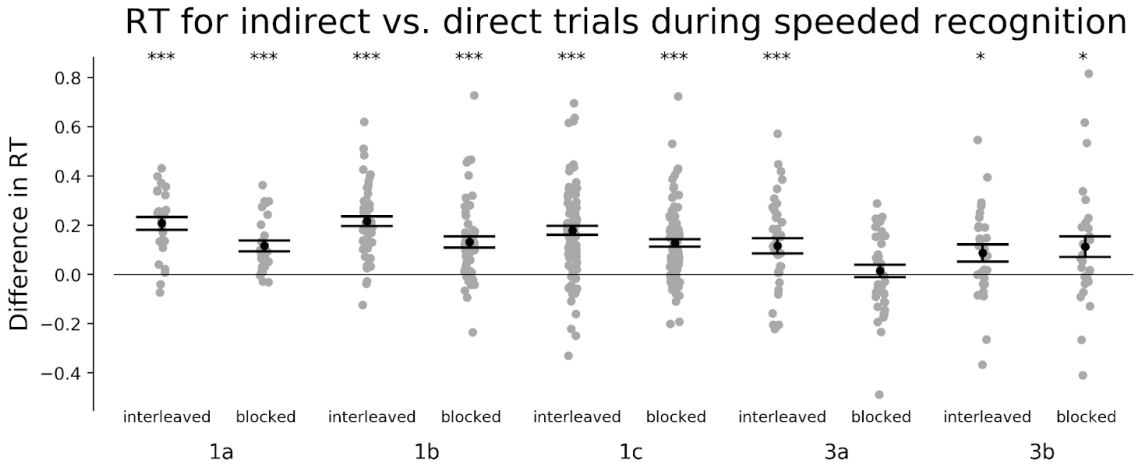

Figure S8. RTs for indirect — direct trials during speeded recognition in Exps 1a-c and 3a-b.

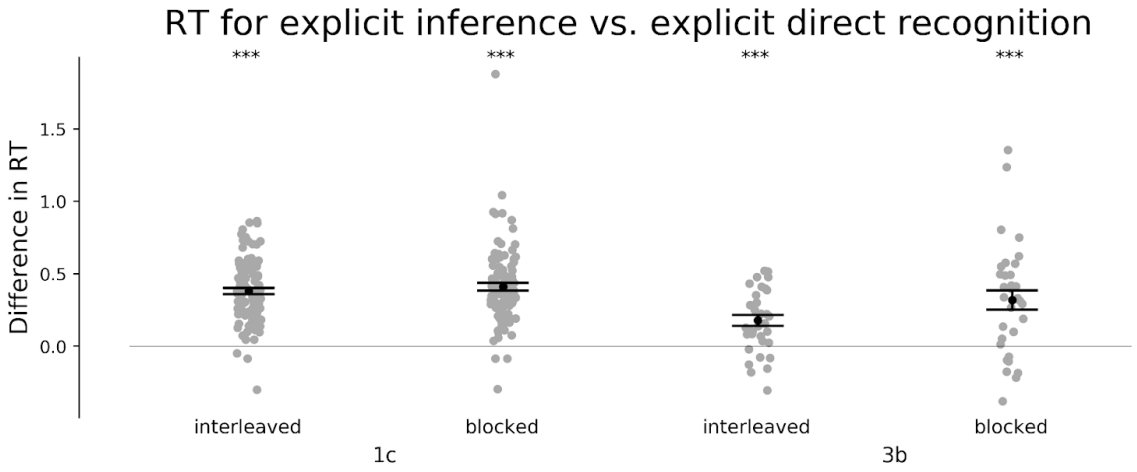

Figure S9. RTs for explicit inference — explicit direct recognition trials by condition in Exps 1c and 3b.

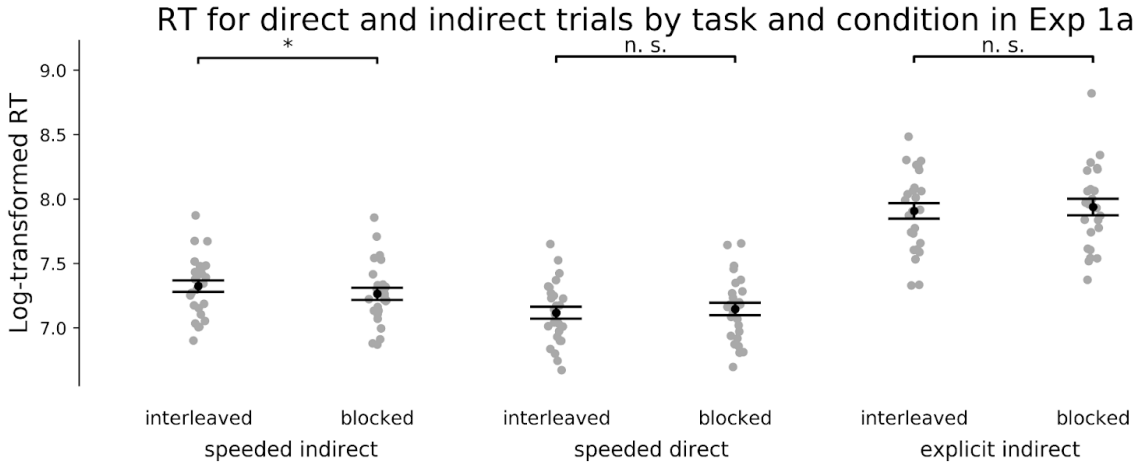

Figure S10. RTs for direct and indirect trials by task and condition in Exp 1a.

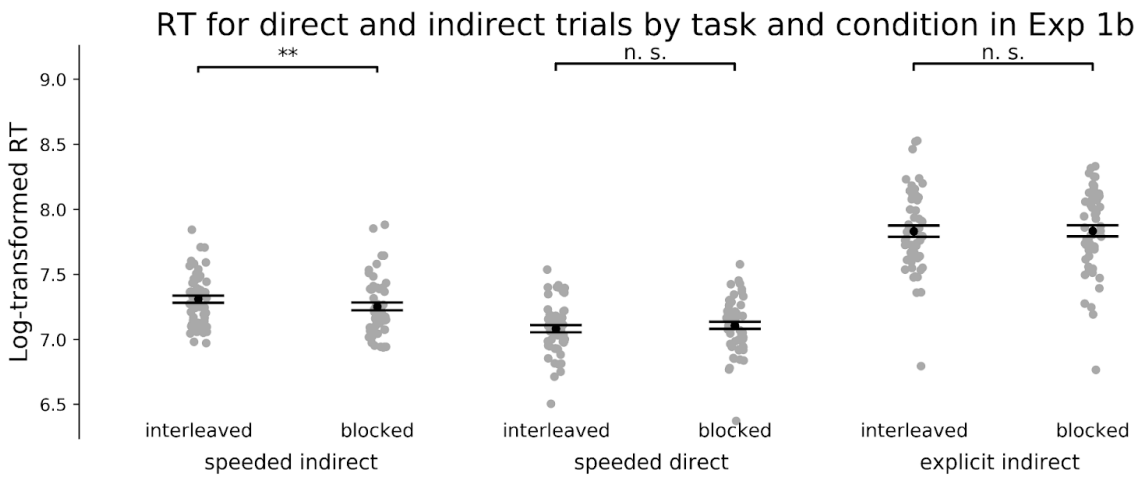

Figure S11. RTs for direct and indirect trials by task and condition in Exp 1b.

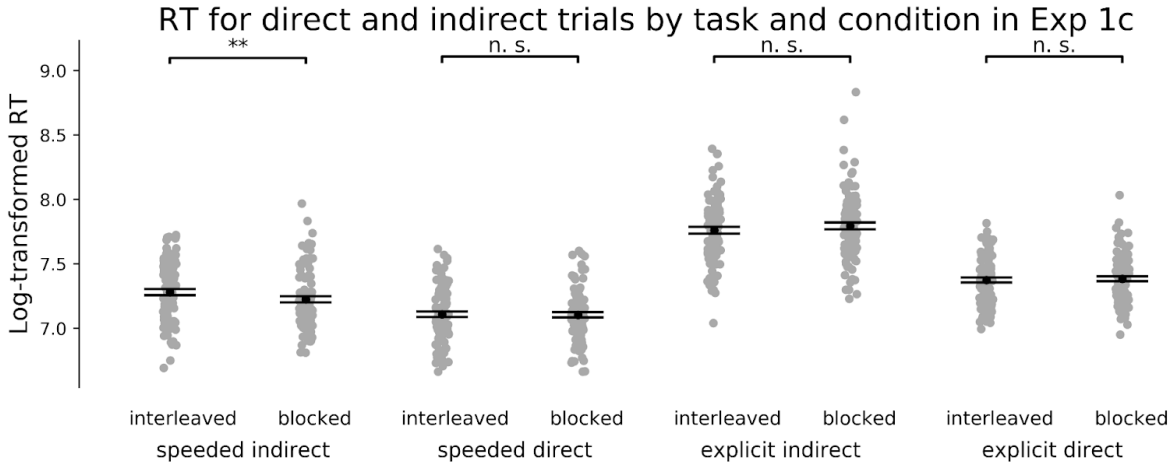

Figure S12. RTs for direct and indirect trials by task and condition in Exp 1c.

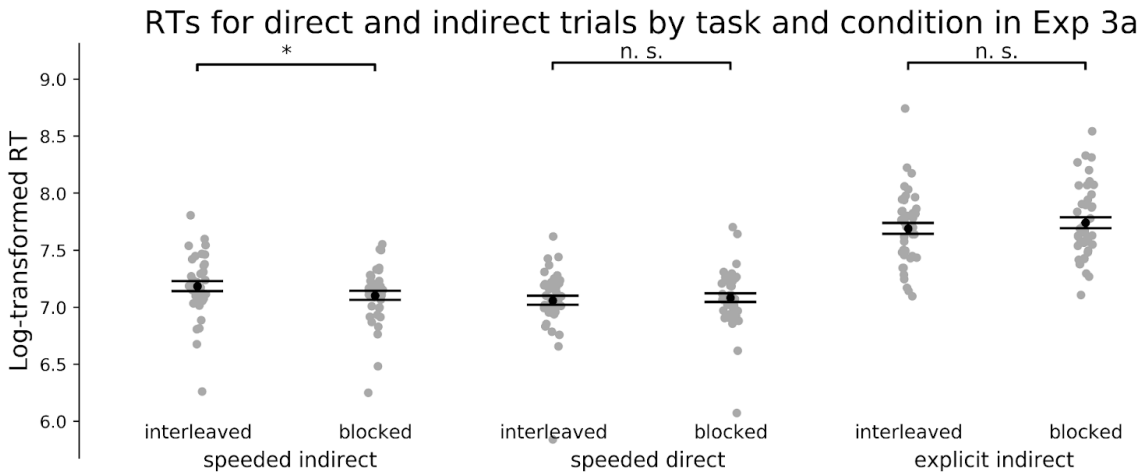

Figure S13. RTs for direct and indirect trials by task and condition in Exp 3a.

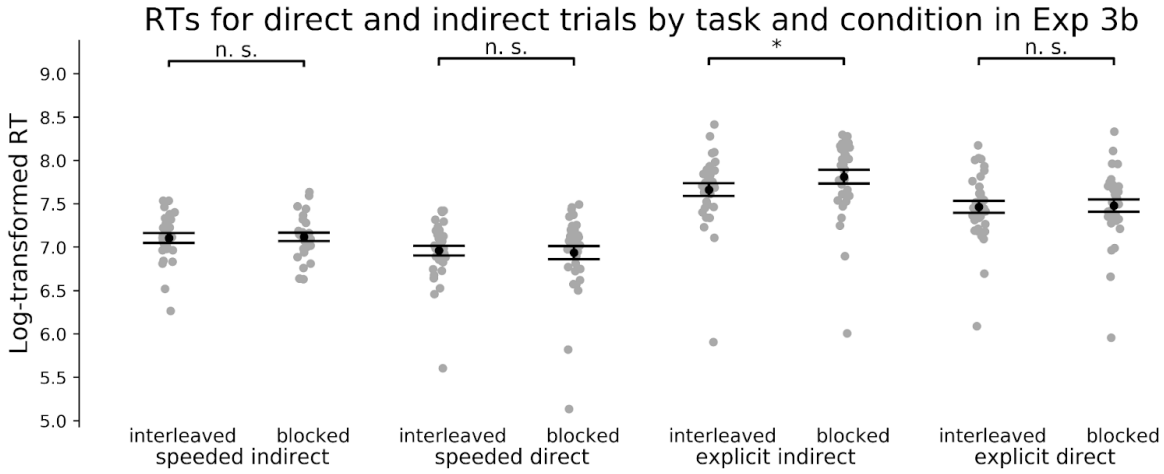

Figure S14. RTs for direct and indirect trials by task and condition in Exp 3b.

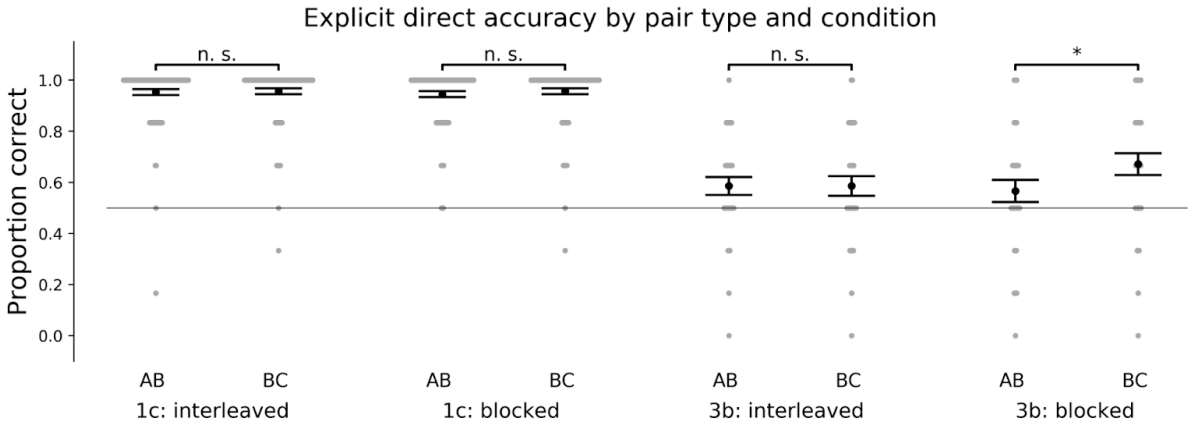

Figure S15. Accuracy for explicit direct trials in Exps 1c and 3b.