Memory & Cognition
2005, 33 (4), 742-750

When logic fails: Implicit transitive inference in humans

MICHAEL J. FRANK and JERRY W. RUDY
University of Colorado, Boulder, Colorado

WILLIAM B. LEVY
University of Virginia, Charlottesville, Virginia

and

RANDALL C. O'REILLY
University of Colorado, Boulder, Colorado

Transitive inference (TI) in animals (e.g., choosing A over C on the basis of knowing that A is better than B and B is better than C) has been interpreted by some as reflecting a declarative logical inference process. We invert this anthropomorphic interpretation by providing evidence that humans can exhibit TI-like behavior on the basis of simpler associative mechanisms that underlie many theories of animal learning. In this study, human participants were trained on a five-pair TI problem (A+B−, B+C−, C+D−, D+E−, E+F−) and, unlike in previous human TI studies, were prevented from becoming explicitly aware of the logical hierarchy, so they could not employ logical reasoning. They were then tested with three problems: B versus D, B versus E, and C versus E. Participants only reliably chose B over E, whereas the other test conditions yielded chance performance. This result is inconsistent with the use of logical reasoning and is instead consistent with an account developed to explain earlier TI studies with rats that found the same pattern of results. In this account, choice performance is based on differential associative strengths across the stimulus items that develop over training, despite equal overt reinforcement.

When told that John is taller than Bill, who is taller than Fred, one would logically infer that John is taller than Fred. This outcome is often referred to as *transitive inference* (TI). It is obvious from the above example that people have the capacity to explicitly encode these statements ("premises"), make the logical inference, and declare the basis of their conclusion. However, this same TI behavior has been demonstrated in a wide variety of animal species (rats, pigeons, and primates; Davis, 1992; Dusek & Eichenbaum, 1997; Van Elzakker, O'Reilly, & Rudy, 2003; von Fersen, Wynne, Delius, & Staddon, 1991; Wynne, 1995). Some researchers have assumed that such animals use a process much like human logical reasoning to achieve these behavioral results (e.g., Davis, 1992; Dusek & Eichenbaum, 1997). Others have argued that these behaviors are better understood as resulting from subtle differences in the associative strength of the stimulus cues (Frank, Rudy, & O'Reilly, 2003; Van Elzakker et al., 2003; von Fersen et al., 1991; Wynne, 1995). In this article, we turn the anthropomorphic bias of the former interpretation on its ear and demonstrate that people can use a nonexplicit means of exhibiting TI-like behavior that has distinguishing characteristics of the associative strength mechanisms. Thus, we conclude that people can be added to the list of species that demonstrate TI-like behavior without relying on explicit logical reasoning.

In animals, TI is evaluated by first training a series of simultaneous discrimination problems (e.g., A+B−, B+C−, C+D−, D+E−), where "+" and "−" refer to the rewarded and nonrewarded choices, respectively. After reaching criterion performance on all training pairs, the animal is tested for inference with novel pairs (e.g., BD and AE). The successful choice of stimulus B over D is taken as evidence of TI. Whereas AE performance is trivial—A is always rewarded and E is never rewarded—the same cannot be said about BD, because B and D are equally often rewarded during training. Some then interpret successful BD performance to indicate that the animal uses relational information to infer that B is logically superior to D (Dusek & Eichenbaum, 1997). However, more recent findings pose a challenge to this account. Specifically, when another premise pair (E+F−) is added to the training paradigm, transitive behavior breaks down in rats (Van Elzakker et al., 2003). When tested, rats reliably chose B over E, but BD performance did not differ statistically from chance.

Whereas a logical inference account predicts similar good performance in both cases, computational model-

We thank Terri Zacharias for testing participants at the University of Virginia. This research was supported by ONR Grants N00014-00-1-0246 and N00014-03-1-0428 and National Institutes of Health Grant MH61316. Correspondence concerning this article should be addressed to R. C. O'Reilly or M. J. Frank, Department of Psychology, University of Colorado, 345 UCB, Boulder, CO 80309 (e-mail: oreilly@psych.colorado.edu or frankmj@psych.colorado.edu).

ing supported an alternative account that is consistent with the observed pattern of results (Frank et al., 2003). The model suggested that differential associative strengths accrue across B and E, despite equal overt reinforcement (Figure 1). Although this finding contradicts the pervasive assumption that equal overt reinforcement implies equal underlying associative strengths, it follows naturally from basic learning mechanisms that were implemented in our model.[1] After training on the TI paradigm, the model developed a net positive association to B and a net negative association to E. This difference was sufficient to induce the model to reliably choose B over E at test. The smaller difference between B and D values explains inferior performance in the BD test case.

The above reasoning suggests that TI-like behavior in animals can be accounted for without invoking the use of logical strategies. Thus, humans should also perform well in TI-like tasks without having to rely on explicit logical reasoning. To test this idea, Greene, Spellman, Dusek, Eichenbaum, and Levy (2001) trained human participants on the four-pair problem [AB, BC, CD, DE], using a training protocol similar to that used in rats. Verbal strategies were limited by using unfamiliar visual stimuli (Japanese hiragana characters). On the BD test, participants successfully chose B over D. It was concluded that explicit awareness was not necessary for TI-like behavior in humans. However, several participants did indeed become aware of the hierarchy as training progressed. Although performance was not correlated with postexperimental measures of awareness, the possibility could not be excluded that some participants were more aware than they reported. Furthermore, the same ordering of hiragana characters was used for each participant, so that BD performance could potentially be attributed to B having more distinct surface features than D. More importantly, the four-pair problem permitted only one novel test pair (BD), making it difficult to ascertain whether performance was dictated by logical reasoning or by differences in associative strength.

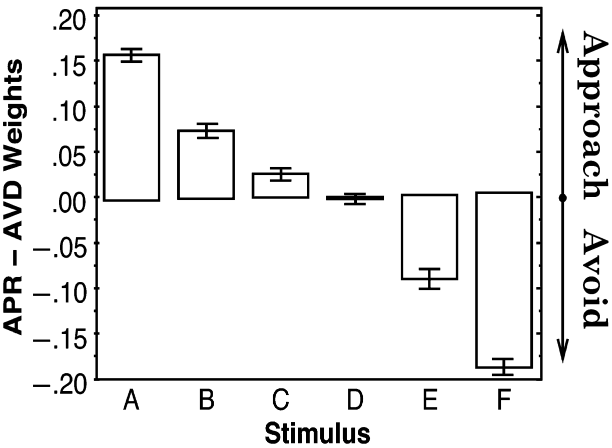


Figure 1. Averaged final weights for the computational model, modified from Frank et al. (2003). The weights shown reflect the difference between the model's approach (APR) and avoid (AVD) associations for each stimulus element. Despite receiving equal overt reinforcement, the model develops a net positive association to B and a net negative association to E, explaining its tendency to correctly choose B over E at test. The smaller difference between B and D associations resulted in unreliable BD performance.

The purpose of the present experiment is to provide a more definitive test of (1) the general proposition that adult humans can display TI-like behavior in the absence of explicit reasoning and (2) the specific hypothesis that such behavior is mediated by an underlying associative structure. To do this, we followed the general experimental paradigm laid out by Greene et al. (2001) but took special measures to reduce the probability that the participants would become explicitly aware of the hierarchy. We also trained them on a five-problem discrimination set (A+B−, B+C−, C+D−, D+E−, E+F−), which allowed us to test them with several novel combinations (BD, CE, BE, and AF). If we were successful in preventing people from using their explicit reasoning skills, the associative account would have predicted a graded outcome, with the strongest evidence of transitive behavior being observed on the AF test and the weakest evidence on the BD test. Conversely, if participants explicitly detected the hierarchy, performance should have been equal and robust across all test pairs.

METHOD

Participants

The participants were 63 undergraduate students from the University of Colorado and 9 undergraduate students from the University of Virginia, participating for course credit.

Stimuli

The stimulus items were six characters selected from the Japanese hiragana script, as in Greene et al. (2001). The assignment of hiragana characters to hierarchical elements A–F was randomized across participants (Figure 2 shows one example of a stimulus hierarchy). The characters were presented on a 19-in. color monitor in 36-point font size.

Procedure

Prior to training, instructions presented on the computer were as follows: "Two black figures will appear simultaneously on the computer screen. You are to select the 'correct' figure as quickly and accurately as possible. At first this will be by trial and error, and you may be confused. Don't worry, you'll have plenty of practice! You will soon find the correct figure is easily learned." No instructions were given that would lead the participant to believe that the stimuli were ordered hierarchically.

For each pair of characters, the participant had to press the "z" key to choose the stimulus on the left or the "m" key to choose the stimulus on the right. The actual location of each individual character (left or right) was counterbalanced across trials, so that no spatial position biases would arise. Feedback was provided to the participant with the word CORRECT! written in blue letters or the word INCORRECT written in red letters. These were the same methods used by Greene et al. (2001).

Training. Training consisted of four phases of blocked trials, followed by a fifth phase of randomly interleaved trials. Each phase was terminated after criterion performance of at least 75% correct on all pairs, and at least 60% on each individual pair, was achieved. In Phase 1, the premise pairs were presented in blocks of six trials, such that the first block consisted of mostly AB trials, the second

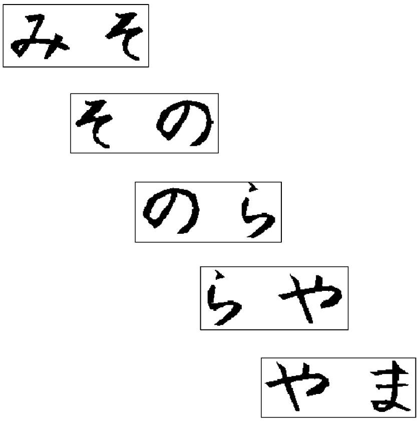


Figure 2. The five pairs of Japanese hiragana stimuli used in the experiment. Each pair was presented separately in different trials, with participants pressing the "z" key to choose the stimulus on the left or the "m" key to choose the stimulus on the right. The hierarchy goes from top to bottom, where the top pair is AB and the bottom pair is EF. In this example, the correct choice is always the stimulus on the left. Note that in actuality, the position of the correct stimulus was randomized across trials, and the assignment of hiragana character to hierarchical element was randomized across participants.

block consisted of mostly BC trials, and so on. However, pilot work demonstrated that if these blocks were completely homogeneous, the participants tended to become aware of the stimulus hierarchy, because of the natural progression from AB to BC to CD, and so forth. To prevent this, "distractor" trials were inserted into a minority of trials in and between each block (see Figure 3). These trials were meant to disrupt the descending order of hierarchical presentation, making the stimulus hierarchy less obvious. Nevertheless, because distractor trials were composed of stimulus combinations from other blocks, they were valid training trials and were therefore included in the analyses for determining whether criterion was met.

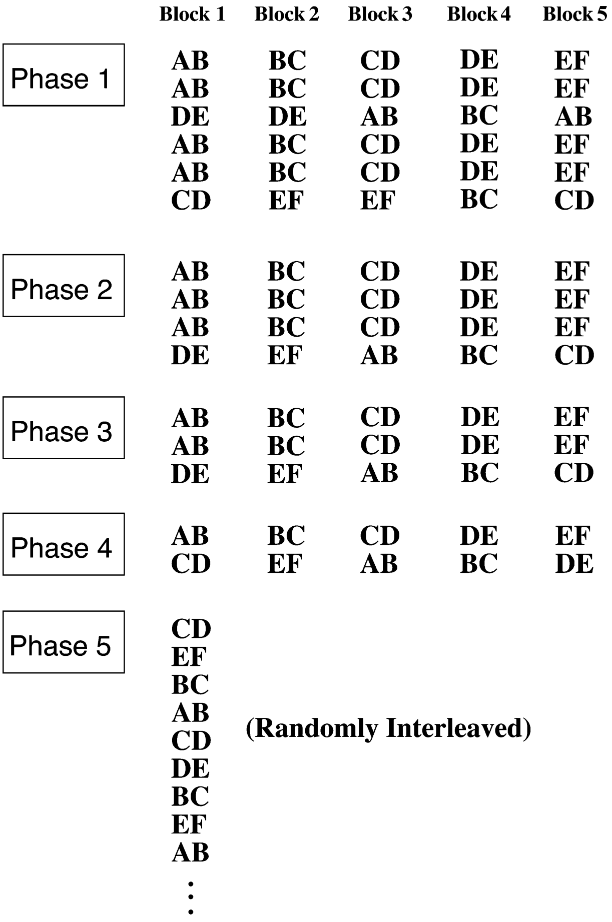
In Phases 2–4, the number of trials per block was decreased, similar to procedures used in rat experiments. See Figure 3 for the number of trials in each block and the actual training sequences. In Phase 5, all pairs were randomly interleaved for a total of 25 trials before criterion performance was evaluated. If criterion was not met, the random sequence was repeated. The purpose of this procedure was to put all participants at essentially the same performance level at test. If the participants failed to meet criterion after several sequence repetitions, they did not continue on to the test phase.

Testing. The test phase was similar to the fifth training phase in that all pairs were randomly interleaved. However, no feedback was provided, and the four transitive pairs BD, BE, CE, and AF were added to the mix of randomly ordered pairs. All pairs were presented six times each.

Postexperimental awareness measures. Different strategies may be used to learn the premise pairs. For example, participants can memorize specific instances of the training stimuli, or they can abstract a general rule. This distinction is particularly relevant for the present study: Instance learning should not be as amenable for logical reasoning, compared with rule abstraction. We therefore needed a measure to help disentangle these two strategies, in order to know whether the participants employed logical reasoning during test.

Before elaborating on our particular measures of awareness, we emphasize that there are at least two possible connotations of "awareness." The first and most straightforward meaning is the degree to which participants are aware of their basis for choosing in individual instances during training (e.g., did they have a "rule" for each response, or did they make choices without explicitly knowing why?). This type of awareness is notoriously difficult to assess using questionnaires—participants may develop some subtle or arbitrary rule undetectable by questionnaires (e.g., Shanks & St. John, 1994). However, this type of awareness is not sufficient for making novel inferences. We were more interested in a higher order measure of awareness that would allow the participants to use explicit inferential reasoning processes during test. Thus we simply had to measure the degree to which the participants were aware of the hierarchical relationship *among* the stimuli; we argue that this higher order awareness is more feasibly detected.

Following the experiment, all participants were given a questionnaire to assess their awareness of the logical hierarchy of the stimuli and to determine what strategies, if any, were used to respond to the novel test pairs. We argue that, at least in this regard, our questionnaire was appropriate since it asked increasingly leading questions to determine whether participants even had the required knowledge that would permit making logical inferences. Furthermore, as is shown below, our measures of awareness had a



	Block 1	Block 2	Block 3	Block 4	Block 5
Phase 1	AB AB DE AB AB CD	BC BC DE BC BC EF	CD CD AB CD CD EF	DE DE BC DE DE BC	EF EF AB EF EF CD
Phase 2	AB AB AB DE	BC BC BC EF	CD CD CD AB	DE DE DE BC	EF EF EF CD
Phase 3	AB AB DE	BC BC EF	CD CD AB	DE DE BC	EF EF CD
Phase 4	AB CD	BC EF	CD AB	DE BC	EF DE
Phase 5	CD EF BC AB CD DE BC EF AB ⋮	(Randomly Interleaved)			

Figure 3. The five training phases of the experiment. In each of the first four phases, stimuli were presented in sequential blocks of trials of decreasing length. Note that the blocks are not completely homogeneous with respect to trial type. The figure depicts the actual number of trials per block in each phase and shows where distractor trials were placed. In Phase 5, the stimulus pairs were randomly interleaved (for 25 trials), as in rat experiments. Participants had to meet a performance criterion of 75% on each phase, and at least 60% on each pair within each phase, before advancing to the next phase.

strong relationship with logical test performance. See the Appendix for details of the questionnaire and typical responses made in the unaware and aware groups.

RESULTS

Of the 72 individuals who participated, 7 failed to meet criterion in Phase 4 or 5 of training and did not advance to the test phase. When we analyzed postexperimental measures of awareness (prior to analyzing test data), another 8 participants were determined to be explicitly aware of stimulus hierarchy and used this knowledge as a rule for choosing among stimuli and test pairs. These participants scored close to 100% on all test pairs, as would be predicted by rational application of hierarchical rules (Figure 4).

The data from the remaining 57 participants who met criterion performance on all training pairs, yet had no explicit knowledge of the hierarchical relationship among the pairs (and thus could not apply explicit logical reasoning to determine the correct choices during the test phase), are analyzed further below.

Premise Pair Performance

Figure 5 shows mean performance on the premise pairs during the final phase of training when premises were interleaved. Note that performance on the anchor premises (AB and EF) was better than that on the internal premises BC, CD, and DE. A repeated measures analysis of variance was applied. An overall accuracy effect of training pair condition was present, showing differences in performance among the different pairs [F(4,224) = 10.69, MS_e = 0.224, p = .0001]. A planned comparison indicated that performance on the anchor pairs AB and EF was significantly better than on the others [F(1,56) = 41.59, MS_e = 0.35, p = .0001]. We discuss the significance of these anchoring effects later in the article.

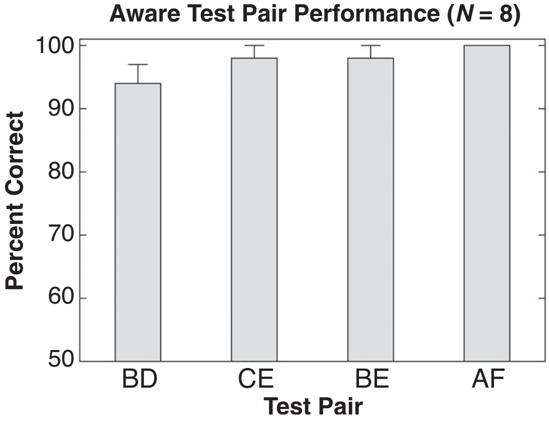


Figure 4. Test pair performance for the 8 participants who became aware of the logical hierarchy. The ceiling effects observed are presumably due to the ability of these participants to correctly apply logical rules in determining their choices at test. Note that the determination of awareness was based on postexperimental questionnaires (see the Appendix) and was blind to the participants' performance. Error bars show standard errors of the mean.

Test Pair Performance

The results of the transitivity tests revealed an overall accuracy effect of test pair condition [F(3,168) = 21.16, MS_e = 1.583, p = .0001] (Figure 6). Although the participants lacked awareness of the stimulus hierarchy, accuracy on BE test trials was significantly better than chance [F(1,56) = 15.84, MS_e = 0.113, p = .0002]. In contrast, performance did not differ from chance for BD [F(1,56) = 1.51, MS_e = 0.123, p = .22]. There was a trend for CE performance to be better than chance, but it did not reach significance [F(1,56) = 3.24, MS_e = 0.108, p = .077]. Planned comparison analysis demonstrated that BE performance was significantly better than both BD [F(1,56) = 5.15, MS_e = 0.819, p = .027] and CE [F(1,56) = 4.28, MS_e = 0.557, p = .043] performance. BD and CE performance did not differ from each other [F(1,56) = 0.12, MS_e = 0.025]. The AF test involved an always-reinforced A stimulus, and a never-reinforced F. Thus, as expected, performance was near ceiling.

Pearson correlation measures were also computed, to determine the possible contributions of individual stimulus elements, as predicted by the associative strength account. There were significant correlations between performance on pairs sharing stimulus elements. Performance on BE correlated with performance on BD (r = .33, p = .013) and CE (r = .41, p = .0015). BE was also correlated with training pair DE (r = .29, p = .03) but not with BC (r = .05, p = .7), indicating that BE performance was dictated more by negative association to E than by positive association to B. Consistent with this conclusion, there was a trend for CE performance to be correlated with training pair DE (r = .24, p = .067), but not with CD (r = .11, p = .41), again pointing to a negative E association. Finally, BD and CE do not share elements, and performance in these two cases was not significantly correlated (r = .097, p = .47).

DISCUSSION

This experiment provides strong evidence that adult humans can respond transitively in the absence of conscious awareness of hierarchical relationships, consistent with earlier findings by Greene et al. (2001). We argue that in the absence of explicit logical reasoning, implicit associative learning processes cause the training elements to acquire different associative strengths, which is sufficient to induce transitive responding. Several lines of evidence support this conclusion. First, transitive choice was displayed by people who had no explicit knowledge of a hierarchical relationship among the test elements. Thus, it is not clear how this behavior could be supported by explicit logical reasoning. Second, this transitive performance was only clearly manifested on the BE test problem and was significantly less evident

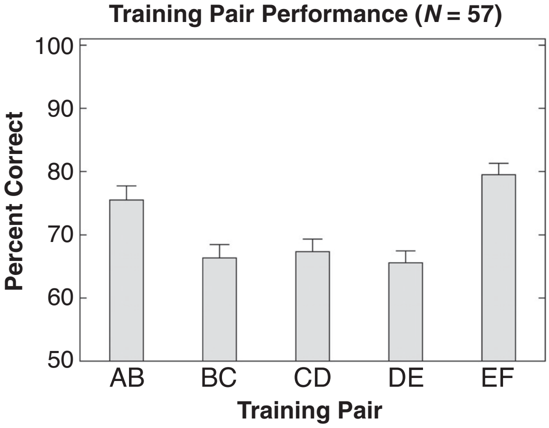


Figure 5. Anchoring effects for premise pairs during interleaved training. Stimuli A and F are the anchors because A is always correct and F is always incorrect. Error bars show standard errors of the mean.

on the BD and CE tests. If the participants had used an explicit inferential reasoning strategy, they would have performed equally well on all test problems. Indeed, the small set of participants who did explicitly detect the hierarchy (A>B>C>D>E>F) displayed equivalent and robust transitivity on all test pairs. Finally, performance on any individual test pair could be reliably predicted by how the participant performed on other pairs that shared a critical element. For example, performance on test pairs BD and CE was correlated with that of BE, since they both share an element with it. An important implication of this reasoning is that there should be no correlation between performance on pairs that did not share an element (e.g., BD and CE), and indeed this was the case.

Awareness Is Not Necessary, but Is Beneficial for Transitive Choice

Prior studies have shown that transitive responding in humans does not depend on explicit awareness (Greene et al., 2001; Siemann & Delius, 1993, 1996). Indeed, these studies found that although some participants became explicitly aware of the logical hierarchy, this did not benefit their choice behavior (relative to unaware participants) during test. That is, unaware participants who were "left in the dark" about the logical hierarchy still exhibited reliable and robust transitivity on the test pairs. Thus, although the basis for transitive responding may seem like logical reasoning by description, its underlying process may instead involve the kind of primitive abilities that allow pigeons and rats to respond transitively in similar paradigms (Markovits & Dumas, 1992). One might point out that young children, who have impoverished conceptual understanding of hierarchical information, fail to respond transitively (Piaget, 1921). However, Trabasso (1975, 1977) showed that these children do in fact respond transitively if they are trained more extensively on the premise pairs, apparently reflecting the fact that their initial performance decrements stemmed from a lack of memory (Bryant & Trabasso, 1971). Trabasso further implied that transitive choice does not recruit logical processes per se but involves the development of a spatial representation of linear order based on differential associative strengths of elemental items (Trabasso & Riley, 1975).

We have shown that unaware participants respond transitively on the BE test pair and therefore maintain that TI-like behavior does not depend exclusively on logical processes. Rather, it can arise from simpler associative processes, in accordance with Wynne (1998) and Delius and Siemann (1998). However, unlike results of previous studies, our results show a clear advantage for transitive responding in the minority of participants who were explicitly aware of the logical hierarchy. These participants employed logical reasoning and performed at near-ceiling levels on all test pairs. We take this to indicate that associative processes are somewhat subtle in that they must compete with participants' biases toward surface features of the particular stimuli and therefore do not lead to perfect TI-like behavior. In contrast, higher order logical reasoning processes available to adult humans—but probably not to animals or young children—reliably lead to transitive choice.

Associative Framework for Unaware Performance

Our account of performance in unaware participants critically depends on the test elements having differential excitatory associative strength, which was predicted by our computational model (Frank et al., 2003). The key question is, why should the test stimuli have different levels of associative strength, given that they were equally often reinforced in the training phases? In brief, our associative strength framework hinges on the role of the

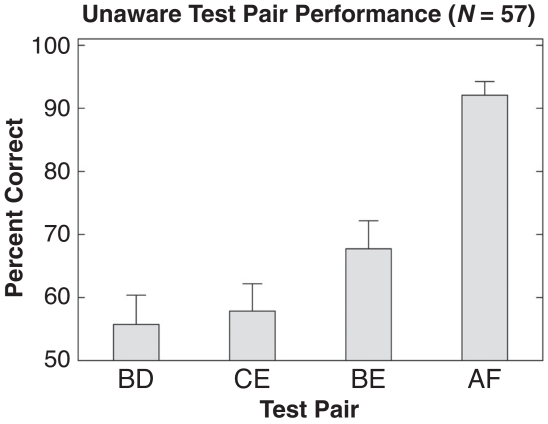


Figure 6. Unaware test pair performance. Choice on test pairs BD and CE did not differ from chance (50%). BE performance was significantly better than performance for both BD and CE. AF is a trivial discrimination, since A is always correct and F is always incorrect. This same pattern of results was observed in transitive performance in rats using an analogous procedure (Van Elzakker et al., 2003). Error bars show standard errors of the mean.

anchor pairs AB and EF in establishing a gradient of associative strength across the elemental stimuli during training. We assume that participants simply learn that A is always correct and F is never correct, and thus do not have to learn anything about the companion stimuli (B in AB, and E in EF). In essence, this is a *blocking effect* (Kamin, 1968), because learning about the companion stimuli is blocked by the perfect predictive reliability of the anchor stimuli. In the case of B, its overall associative strength can then increase to facilitate performance on BC trials, whereas that of E decreases in DE trials. The resulting difference in associative values explains why B is chosen over E in this study with humans and prior studies with animals. In other words, this *anchoring effect* causes approach-and-avoid associations to "bleed" over to adjacent stimuli B and E (for more explicit analyses of these and related ideas, see Frank et al., 2003; Levy & Wu, 1997; Siemann & Delius, 1998; cf. von Fersen et al., 1991).

Our associative interpretation of TI-like behavior in rats and in people unaware of the stimulus hierarchy contrasts with the relational account of Dusek and Eichenbaum (1997). That account assumes equal underlying associative strengths among stimulus elements and further assumes that TI-like behavior is produced by mechanisms (in the hippocampus) that detect an ordered hierarchy and flexibly relates premise pairs during the transitive test. The exact mechanisms by which the hierarchy is established and detected remain unspecified. Unless this account is qualified, it is unclear why it would not predict that unaware participants display equal TI performance on all the relevant test pairs of our experiment (BE, CE, and BD).

Symbolic Distance Effects

One possible explanation for the present data would be to appeal to the symbolic distance (SD) effect (e.g., D'Amato & Colombo, 1990; Hamilton & Sanford, 1978), where it is easier to distinguish symbols that are farther apart on the continuum (i.e., BE is easier to discriminate than is BD, explaining the observed pattern of results). However, in our view, the symbolic distance effect is more of a description of the data than a mechanistic theory. In many ways, our associative account *explains* the observed SD effect, in that we posit that items having greater associative strength differences will result in stronger choice preferences for the more positive item. Indeed, one might label this account the associative distance (AD) model. We point out that lacking a clear, mechanistic explanation for what establishes a symbolic hierarchy in the mind of the participant, one must assume the presence of such a hierarchy or continuum in somewhat unspecified symbolic terms. In contrast, our account provides a concrete, mechanistic explanation for the origin and nature of the underlying differences between items, in terms of differential associative strengths.

Furthermore, as we argued in Van Elzakker et al. (2003), the SD effect is not observed in the BD test case for the four-pair version of the TI task (e.g., Dusek & Eichenbaum, 1997), where BD performance was essentially equivalent to that of the BE problem in the five-pair version (using the same experimental paradigm across both versions, in rats; Van Elzakker et al., 2003). The associative account, in contrast, readily explains this pattern of results in terms of the test item's proximity to the anchors (i.e., BD in the four-pair version are each one item away from the end items of A and E, as are BE in the five-pair version). See Van Elzakker et al. for more discussion of these issues and alternative (e.g., normalized) versions of SD that also do not explain this data very well.

Here, we provide several additional lines of evidence of our associative account that go beyond simple SD effects. First, unless specified otherwise, the SD account assumes that each item is arranged with uniform spacing along a continuum of strength. In contrast to this uniform spacing, the learning mechanisms in our computational model produce an uneven distribution of associative strengths, which are consistent with the "flat U" shape of the training item performance (see Figure 5). In other words, items closer to the anchor points have greater associative strength differences than those in the middle.

Second, the SD effect predicts equal BD and CE performance, because each pair is separated by one symbol. In contrast, our model predicted that CE performance should be intermediate between BD and BE. This prediction stems from the fact that in the model (and in rats), the anchoring effects were asymmetrical, such that EF performance was greater than that of AB. Thus, E had more of a negative associative value than B had a positive value (Figure 1), thereby making CE performance better than that of BD. However, because anchoring effects in the present study were more or less symmetrical—EF performance was only slightly greater than that of AB—we could not be as confident about our CE>BD prediction. Indeed, CE performance was only numerically but not significantly better than that of BD. However, it should be noted that the BE>CE>BD pattern of results has been found in other nonverbal TI tasks (Potts, 1977; Werner, Koppl, & Delius, 1992). In the one study showing an anchoring effect that was asymmetrical in the direction opposite what is usually observed (i.e., AB>EF), the BD/CE difference was also in the opposite direction (BD>CE; Siemann & Delius, 1993), as predicted by our associative account. Further research is needed to determine why and under what conditions AB performance should be better or worse than EF performance—in other words, under which conditions is it easier to learn that one stimulus (A) is good compared with learning that another stimulus (F) is bad?

Third, correlational analysis in our study demonstrated that performance on any individual pair could be predicted by the participant's tendency to perform well on other pairs sharing a critical stimulus element. This analysis strongly suggests that performance is dictated by elemental associative strength.

We have provided several analyses that (indirectly) provide support for the associative strength hypothesis that goes beyond simply pointing out standard SD effects. More direct testing of the associative hypothesis may come from further studies in which it predicts a lack of SD effect. For example, by employing an extended six-pair version of the task (A>B>C>D>E>F>G), several test conditions are possible. Note that test pairs BD, CE, and DF are all separated by one symbol. Whereas the SD effect predicts equal performance in all these cases, our associative strength hypothesis predicts that BD and DF are above chance (because each involves a stimulus adjacent to an anchor element) but that CE performance should be significantly worse.

Neural Correlates of Transitive Choice

The hippocampal formation has been implicated as playing an important role in TI behavior, in large part from the finding that damage to it disrupts TI-like behavior in rats (Dusek & Eichenbaum, 1997). These authors argued that this finding is consistent with the more general view of the relational account that the hippocampus is involved in explicit, declarative memory processes. In this context, our finding of implicit TI-like behavior in unaware humans that is very similar to that of intact rats may be surprising; one might have predicted that unaware humans would perform more like hippocampally lesioned rats. However, our animal studies, together with our computational modeling work (Frank et al., 2003; Van Elzakker et al., 2003) have led us to believe that the hippocampus makes a relatively minor contribution to TI-like behavior in rats and unaware people. Our model predicts that the hippocampus is important only for the early phases of learning the premise pairs and does not play an active role during behavior on the test pairs. Specifically, in our combined hippocampal–cortical model, the hippocampus rapidly encodes the distinct anchor pairs as separate conjunctions and therefore internally blocks learning about the positive aspect of E in EF trials and the negative aspect of B in AB trials. This blocking effect, which is exaggerated relative to the purely cortical model, led to a stronger net positive association for B and a stronger net negative association for E. These associative weights then produced stronger TI-like behavior at test than was observed in the purely cortical model. However, as the models were trained longer, slower developing elemental blocking signals in the cortex (i.e., as the animal learns to choose A, it no longer learns negative aspects of B) produced similarly strong differential associative weights to B and E, to the point that transitive choice no longer depended on the hippocampus. We therefore posit that the hippocampus is not fundamentally critical for developing differences in associative strength required for transitive responding, although it helps them to develop faster.

Neuroimaging methods were used to assess activity in the hippocampus in participants who were instructed to look for a hierarchical relationship among the stimuli and were therefore employing logical reasoning on the test for transitivity (Nagode & Pardo, 2002). Notably, hippocampal activation was observed only in the early stages of training on the premises and not during testing. This result is generally consistent with our computational model. More importantly, it shows that the hippocampus is not preferentially active while participants make inferences, as implied by the relational account (Dusek & Eichenbaum, 1997). On the other hand, the prefrontal and parietal cortices, which are known to be engaged in higher level cognitive functions, were activated in another explicit TI task, as one might expect (Acuna, Eliassen, Donoghue, & Sanes, 2002). Nevertheless, a more recent study showed hippocampal activation during explicit transitive judgments (Heckers, Zalesak, Weiss, Ditman, & Titone, 2004). Again, we emphasize that explicit forms of TI that require awareness of hierarchical structure may qualitatively differ in underlying psychological and neural processes. In the present study, aware subjects exhibited uniformly good TI performance across all test pairs and thus did not likely operate purely on the basis of differential associative strengths. Hippocampal pattern completion mechanisms probably play some role in this explicit form of TI performance (Levy & Wu, 1997; O'Reilly & Rudy, 2001).

CONCLUSION

We have provided evidence that people can display transitivity even when they are unaware of the hierarchical ordering among the test stimuli. However, transitive behavior was selective to just one (BE) of the three relevant test combinations (BE, CE, and BD). In contrast, the few participants who became aware of the ordering displayed robust transitivity to all test pairs. This pattern of data falls naturally out of the associative account, where performance is based on the relative associative strengths of the different items, as established during training. Furthermore, it may be harder to explain in other frameworks that postulate a more regular logical or symbolic ordering of items. Thus, we conclude that relatively simple associative processes provide the basis for transitivity when logic is not available.

REFERENCES

Acuna, B., Eliassen, J., Donoghue, J., & Sanes, J. (2002). Frontal and parietal lobe activation during transitive inference in humans. *Cerebral Cortex*, **12**, 1312-1321.

Bryant, P., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, **232**, 456-458.

D'Amato, M. R., & Colombo, M. (1990). The symbolic distance effect in monkeys (*Cebus apella*). *Animal Learning & Behavior*, **18**, 133-140.

Davis, H. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology*, **106**, 342-349.

Delius, J. D., & Siemann, M. (1998). Transitive responding in animals and humans: Exaptation rather than adaptation? *Behavioural Processes*, **42**, 107-137.

Dusek, J. A., & Eichenbaum, H. [B.] (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences*, **94**, 7109-7114.

Frank, M., Rudy, J., & O'Reilly, R. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus: II. A computational analysis. *Hippocampus*, **13**, 341-354.

Greene, A. J., Spellman, B. A., Dusek, J. A., Eichenbaum, H. B., & Levy, W. B. (2001). Relational learning with and without awareness:

Transitive inference using nonverbal stimuli in humans. *Memory & Cognition*, **29**, 893-902.

Hamilton, J. M., & Sanford, A. J. (1978). The symbolic distance effect for alphabetic order judgements: A subjective report and reaction time analysis. *Quarterly Journal of Experimental Psychology*, **30**, 33-41.

Heckers, S., Zalesak, M., Weiss, A., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, **14**, 153-162.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive Stimulation* (pp. 9-31). Miami: University of Miami Press.

Levy, W. B., & Wu, X. (1997). A simple, biologically motivated neural network solves the transitive inference problem. *Proceedings of the IEEE International Conference on Neural Networks* (Vol. 1, pp. 368-371). Piscataway, NJ: IEEE Press.

Markovits, H., & Dumas, C. (1992). Can pigeons really make transitive inferences? *Journal of Experimental Psychology: Animal Behavior Processes*, **18**, 311-312.

Nagode, J. C., & Pardo, J. V. (2002). Human hippocampal activation during transitive inference. *NeuroReport*, **13**, 939-944.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, **108**, 311-345.

Piaget, J. (1921). Une forme verbale de la comparaison chez l'enfant [A verbal form of comparisons in the child]. *Archives de Psychologie*, **18**, 141-172.

Potts, G. (1977). Frequency information and distance effects: A reply to Humphreys. *Journal of Verbal Learning & Verbal Behavior*, **16**, 479-487.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral & Brain Sciences*, **17**, 367-448.

Siemann, M., & Delius, J. D. (1993). Implicit deductive reasoning in humans. *Naturwissenschaften*, **80**, 364-366.

Siemann, M., & Delius, J. D. (1996). Influences of task concreteness upon transitive responding in humans. *Psychological Research*, **59**, 81-93.

Siemann, M., & Delius, J. D. (1998). Algebraic learning and neural network models for transitive and non-transitive responding. *European Journal of Cognitive Psychology*, **10**, 307-334.

Trabasso, T. (1975). Representation, memory, and reasoning: How do we make transitive inferences? In A. Pick (Ed.), *Minnesota Symposia on Child Psychology* (Vol. 9, pp. 135-172). Minneapolis: University of Minnesota Press.

Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R. Kail & J. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 333-366). Hillsdale, NJ: Erlbaum.

Trabasso, T., & Riley, C. (1975). On the construction and use of representations involving linear order. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 381-410). Hillsdale, NJ: Erlbaum.

Van Elzakker, M., O'Reilly, R. C., & Rudy, J. W. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus: I. An empirical analysis. *Hippocampus*, **13**, 334-340.

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

Werner, U. B., Koppl, U., & Delius, J. D. (1992). Transitive inferenz bei nicht-verbaler Aufgabendarbeitung [Transitive inference in nonverbal task presentation]. *Zeitschrift für experimentelle und angewandte Psychologie*, **39**, 662-683.

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

Wynne, C. (1998). A minimal model of transitive inference. In C. Wynne & J. Staddon (Eds.), *Models of action* (pp. 269-307). Mahwah, NJ: Erlbaum.

NOTE

1. According to this view, the novel test pairs do not satisfy the requirement that the choice cues have equal associative strength, and therefore they do not provide a true test of TI (Delius & Siemann, 1998). Any attempt to make a strong case for a mechanism other than differential associative strengths would require a means of independently verifying the equality of the underlying associative strengths.

(*Continued on next page*)

APPENDIX

Postexperiment Questionnaire Analysis

Eight questions were asked, as follows.

1. Do you have any prior knowledge of the symbols used in the experiment?
2. If you answered "Yes" to question 1, please indicate to what extent you are familiar with these characters.
3. Did you have the impression that some of the pairs were easier to choose from than others?
4. Did you think any of the symbols were ALWAYS correct (no matter what the other symbol was)?
5. Did you think any of the symbols were ALWAYS incorrect (no matter what the other symbol was)?
6. Did you have the impression that there was some kind of logical rule, order, or hierarchy of symbols in the experiment? If so, please explain briefly.
7. In the test phase, were there any new symbols or new combinations of symbols?
8. If you answered "Yes" to question 7, how did you make your choice in these cases? (e.g., guessed, went with instinct, used some sort of rule—explain)

Awareness judgments were made by assessing the above written questionnaires and asking participants to clarify some responses, while being completely blind to their performance.

Eight out of 72 participants were judged to be aware of the logical hierarchy ordering. All 8 aware participants noticed that A was always correct and F was always incorrect. These participants described using the logical rule or hierarchy to determine their choice at test, either explicitly stating that the symbols were "like numbers, where one was greater than another, and I just had to compare the two values" or using the typical explicit inference account: "if A was correct over B and B was correct over C, then A must be correct over C." All of these aware participants satisfied training criteria and proceeded to the test phase. Their testing data are analyzed in the Results section above.

The remaining 64 participants were judged to be unaware of any notion of logical order or hierarchy among premise pairs. Twenty-eight out of 64 participants did in fact notice that one stimulus (A) was always correct, and 29 of them noticed that one stimulus (F) was always incorrect. Six of these participants actually noticed both that A was always correct and F was always incorrect but were still unaware of the overall hierarchical structure of the stimuli. When asked to describe the "rule," some of these participants stated that they memorized specific pairs but could not describe any notion of logical order and did not explicitly know how to respond to the novel test pairs because they had not memorized the correct response to them during training. Others stated that they tried to find a rule dictating why each stimulus was correct or incorrect but "no matter what I thought, it didn't fit."

Moreover, the 57 unaware participants who advanced to the test phase did not use any logical rule or order to determine their choices during test. Many did not notice that there were novel test pairs that differed from the training pairs, and those that did simply guessed, or went with "instinct." A few participants employed an explicit rule that was incorrect (e.g., "I chose the symbol that was widest"). Nonetheless, these participants scored significantly better than chance on test pair BE but not on BD or CE. However, the presence of these kinds of incorrect strategies contributed a significant amount of noise, especially with the randomized stimulus orderings, which is why a relatively large number of participants were used in this study.

Verbal Encoding of Stimuli

Additional informal debriefing of participants revealed that virtually everyone utilized some type of naming strategy in order to learn the premise pairs. Many of them translated the hiragana characters to the closest looking letter of the English alphabet. This suggests that verbal encoding strategies were not completely eliminated, which was not unexpected, and which does not detract from our results. The important criterion for implicit inference here is not that the training pairs are learned entirely implicitly but rather that participants are prevented from explicitly encoding the logical relationship *among* the stimuli. Furthermore, given the difficulty that participants had in learning this version of the task, any attempt to further prevent verbal encoding (e.g., the introduction of a dual task) would likely make the task beyond the capability of most of our participants.

(Manuscript received November 26, 2003;
revision accepted for publication July 30, 2004.)