Neural Correlates of Species-typical Illogical Cognitive Bias in Human Inference

Akitoshi Ogawa[1], Yumiko Yamazaki[1,2], Kenichi Ueno[1], Kang Cheng[1], and Atsushi Iriki[1]

Abstract

■ The ability to think logically is a hallmark of human intelligence, yet our innate inferential abilities are marked by implicit biases that often lead to illogical inference. For example, given AB ("if A then B"), people frequently but fallaciously infer the inverse, BA. This mode of inference, called symmetry, is logically invalid because, although it *may* be true, it is not *necessarily* true. Given pairs of conditional relations, such as AB and BC, humans reflexively perform two additional modes of inference: transitivity, whereby one (validly) infers AC; and equivalence, whereby one (invalidly) infers CA. In sharp contrast, nonhuman animals can handle transitivity but can rarely be made to acquire symmetry or equivalence. In the present study, human subjects performed logical and illogical inferences about the relations between abstract, visually presented figures while their brain activation was monitored with fMRI. The prefrontal, medial frontal, and intraparietal cortices were activated during all modes of inference. Additional activation in the precuneus and posterior parietal cortex was observed during transitivity and equivalence, which may reflect the need to retrieve the intermediate stimulus (B) from memory. Surprisingly, the patterns of brain activation in illogical and logical inference were very similar. We conclude that the observed inference-related fronto-parietal network is adapted for processing categorical, but not logical, structures of association among stimuli. Humans might prefer categorization over the memorization of logical structures in order to minimize the cognitive working memory load when processing large volumes of information. ■

INTRODUCTION

The ability to make strictly logical (syllogistic) inferences is one of our species' unique and defining cognitive features. However, in the absence of cultural training and careful effort, our innate reasoning faculties are characterized by implicit biases that robustly predispose us to make illogical inferences. For instance, if one's son is named Michael, one tends automatically to recall one's son whenever the name Michael is heard. More generally, people who have learned the conditional relation AB ("if A then B") readily but fallaciously infer the inverse, BA. This mode of inference, called symmetry, is strictly illogical because even though it *may* be true, it is not *necessarily* true. Additionally, given an interrelated pair of conditional relations such as AB and BC, people readily show the emergent inferences of AC ("transitivity," which is logical; Goel, 2007) and CA ("equivalence," which is illogical; Sidman & Tailby, 1982), as well as the symmetrical relations BA and CB. (Figure 1). (One could argue that symmetry constitutes a simple form of "equivalence," but as we discuss later, the inference of CA is a conceptually distinct process.) The inferences formed via symmetry, transitivity, and equivalence are derived or emergent relations, because none of them are ever explicitly trained and subjects spontaneously make these inferences without feedback.

In sharp contrast to humans, these biases are mostly absent from the behavior and decision-making of other animals. The fact that these modes of inference are species-typical for humankind may reflect an adaptive origin based on their utility in daily life and survival. For example, we all know that "If it was raining, then the ground is wet." If we look out the window and see that the ground is wet, our minds typically leap to the converse—"If the ground is wet, then it was raining"—and infer that it has, in fact, been raining, without considering all the other ways the ground may have become wet. As a result, we decide to take an umbrella with us when we go out. Although we use this "logic" naturally, strictly speaking, it is illogical, meaning not necessarily entailed. Nevertheless, our illogical bias in a case like this—where the risk from misjudgment is low—is useful, efficient, and practical for guiding our actions. The fact that we so readily and reflexively infer symmetry and equivalence relations (Yamazaki, 2004; Sidman & Tailby, 1982) suggests that the underlying mental strategy was not originally adapted for dealing with purely logical structures. Instead, humans may prefer categorizing—forming paired or grouped, omnidirectional chains or webs of association among incidentally related events regardless of their logical directionality. The advantage of this could be to reduce the

[1]RIKEN Brain Science Institute, Wako-shi, Japan, [2]Keio University, Tokyo, Japan

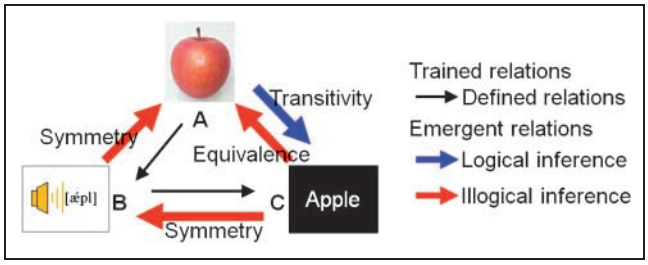

Figure 1. Framework for stimulus equivalence. Given a pair of conditional relations such as AB and BC, people readily and fallaciously infer the inverse, BA and CB (symmetry, illogical). Moreover, the people readily infer both AC (transitivity, logical) and CA (equivalence, illogical).

cognitive load on working memory when thinking about paired or grouped events, and thereby allowing us to process larger volumes of information.

Furthermore, it is worth noting that these human-typical modes of inference may be more than merely useful; it is possible that our innate bias toward symmetric inference is a foundational component of our linguistic faculties of naming and labeling. Again, if one's son is named Michael, then if the name Michael is heard, one tends automatically to recall one's son. Likewise, when a mother points to a picture of an apple and says "apple," her child automatically—but strictly speaking, illogically—infers that the relationship between the utterance and the picture is bidirectional and unifying. This bias, known as a part of stimulus equivalence, emerges concomitant with language development in childhood, following an explosive increase in vocabulary (Dickins & Dickins, 2001). The emergence of stimulus equivalence is retarded in some high-functioning autistic children (Eikeseth & Smith, 1992) and other forms of mental retardation characterized by reduced verbal communication faculties (O'Donnell & Saunders, 2003). However, although the circumstantial evidence is tantalizing, whether there is a deep link between language development and stimulus equivalence remains open and controversial.

Despite their importance, few studies have attempted to identify the brain mechanisms subserving these robust biases in human inference. In part, this is because animal models of these mechanisms are extremely limited in nature and are very difficult to replicate through training. Animals including the rat (Hank, 1992), pigeon (von Fersen, Wynne, Delius, & Staddon, 1991; D'Amato, Salmon, Loukas, & Tomie, 1985), monkey (Treichler & Van Tilburg, 1996; D'Amato et al., 1985), and chimpanzee (Boysen, Berntson, Shreyer, & Quigley, 1993) can handle transitivity, but they can rarely be made to acquire symmetry or equivalence. Among the most notable behavioral evidence, illogical inference that was strikingly similar to that of humans has been reported in California sea lions (Kastak & Schusterman, 2002a, 2002b; Schusterman & Kastak, 1998). In these studies, the animals were trained with two types of initial "if–then" propositions (AB and BC) using graphical figures as stimuli. In follow-up tests, the animals spontaneously expressed all three types of the emergent relations mentioned above (symmetry, transitivity, and equivalence). The only other species that has proven able to partially pass such tests, albeit with extreme difficulty, was a chimpanzee (Tomonaga, Matsuzawa, Fujita, & Yamamoto, 1991). And because invasive neurophysiological studies of higher primates and pinnipeds are ethically fraught, not to mention prohibitively expensive, the only mode of inference to which neuroscientists have had easy access is transitivity in human subjects (Goel, 2007; Goel, Buchel, Frith, & Dolan, 2000).

In humans, the neural correlates of transitive inference have been extensively studied using various neuroimaging techniques. These studies have notably implicated cingulate cortex, basal ganglia, as well as frontal, temporal, occipital, and parietal lobes (Goel, 2007). In contrast, neuroimaging has rarely been employed to probe the neural correlates symmetry and equivalence. To date, only two series of fMRI studies (Schlund, Cataldo, & Hoehn-Saric, 2008; Schlund, Hoehn-Saric, & Cataldo, 2007; Dickins, 2005; Dickins et al., 2001) have looked into this matter. Both studies limited their scanning to the lateral prefrontal and posterior parietal cortices; and because they focused on propositional-linguistic inference, the investigators made their comparisons only in relation to a language-related control task in the left hemisphere. Furthermore, because the stimuli they used (i.e., iconic stimuli of Macintosh Mobile font, ASCII characters, and consonants and vowels) were already familiar to their subjects, their experimental task could have been engaging mental functions that mediate the retrieval of learned categories (Schlund et al., 2007; Dickins et al., 2001) rather than the immediate inference of new categories, for which the neural mechanisms are known to be different (Ashby & Ell, 2001). It was also suggested that stimulus familiarity influenced the performance of transitivity and equivalence tests (Saunders & Green, 1999). If the goal is to isolate the cognitive processes related to inference, it is more appropriate to use unfamiliar stimuli.

Human inferential ability has been investigated primarily using natural-language sentences as stimuli (Goel, 2007; Parsons & Osherson, 2001). In contrast, the inferential abilities of animals have been investigated primarily using visual, auditory, or other sensory stimuli. But if an inference is defined as a spontaneous, untrained derivation from trained stimulus relations, then human inferential ability is likewise investigable using nonlinguistic stimuli. In the present study, human subjects were trained on several pairs of conditional "if–then" relations between unfamiliar visual stimuli that were difficult to name. After five sets of AB and BC relations were very well learned, subjects were tested in the fMRI scanner with the four novel, untrained sets of BA (symmetry), CB (symmetry), AC (transitivity), and CA (equivalence). Because these test relations were novel, the associated activation should correspond purely to the formation and manipulation of the emergent relations

and not to memory retrieval. This procedure allowed direct comparison of the activation patterns associated with all three modes of inference throughout the brain.

METHODS

Subjects

Fifteen healthy right-handed adults (6 women and 9 men; aged 21–35 years, mean = 27.5 years) with no history of neurological or psychiatric disorders participated in the study. The experimental procedure was approved by The Third Research Ethics Committee of RIKEN. Written informed consent was obtained from all subjects prior to the experiments.

Task

Stimuli

Fifteen stimulus figures were generated for the experiment. Each figure consisted of seven white squares arranged in meaningless combinations within a four-by-four grid of black squares. The initial allocation rule was that each row and each column of each figure must contain exactly two white squares and two black squares. As a final step, to create vertical and horizontal asymmetry for each figure, a white square was randomly replaced by a black square. During training and testing, the figures were always presented on a black background, and the average brightness of the white squares was always the same magnitude. Logical relations between the figures were defined by randomly sorting them into three groups of five figures each, labeled A, B, and C (Figure 2A). The grouping of the stimulus figures was randomized differently for each subject. The A, B, and C group labels were used only by the experimenters and were not shared with the subjects.

Matching-to-Sample Procedure

Each subject participated in a total of three experiments: Two behavioral experiments, which were administered in front of a desktop computer, and one fMRI experiment, which was performed inside the scanner. The following protocols were common to both training and test sessions. At the onset of each trial, a fixation point, a small white cross, appeared for 1500 msec at the center of the display. A lone "sample" stimulus figure then replaced the fixation point. After 800 msec, two "comparison" figures appeared to either side of the sample (Figure 2B). The subjects were asked to choose the comparison figure they thought was related to the sample figure, even if they felt uncertain about their selection. They had 2700 msec to choose the left or right comparison by pressing the left or right response button. When the time elapsed, all three figures disappeared from the display and the subjects received feedback. No feedback was delivered if the subjects failed to act in the 2700 msec time limit, whereas auditory feedback (a bell sound for "Correct" or a buzzer sound for "Incorrect") was delivered in the training sessions, and visual feedback (a green "X" or "O" at the center of the display) was delivered during the test session in the fMRI scanner.

Training and Testing

Training Experiments

The first training session was always administered at least one night, but no more than a month, before the test session. During the first training session, the subjects learned five pairs of stimulus–figure relations using the matching-to-sample procedure described above. Each of these pairs contained one AB relation and one BC relation. The stimulus figures and the logical relations between

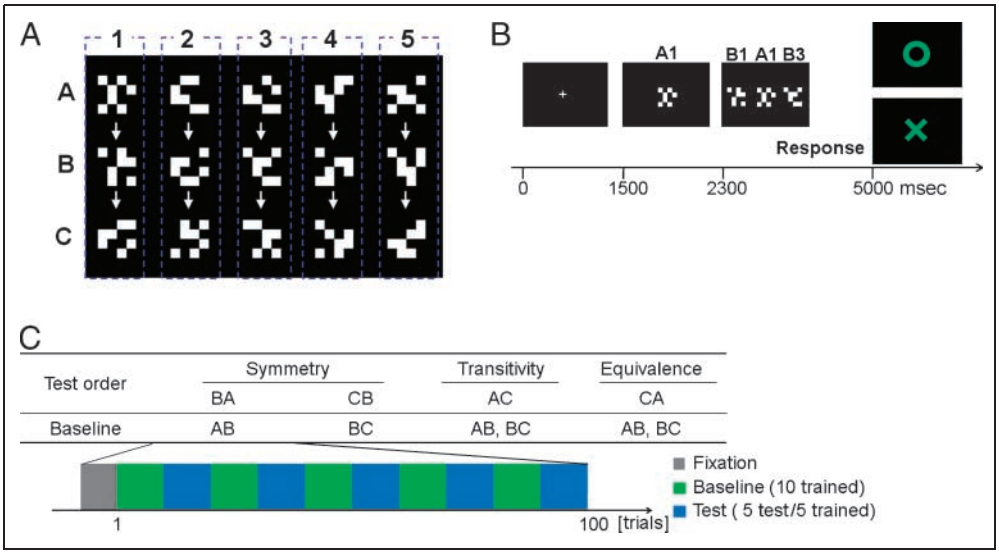

Figure 2. (A) An example of stimulus assignment. Five sets (blue-dot boxes) of the relations of AB and BC were trained. Arrows indicate the defined, unidirectional relations. (B) Matching-to-sample task. The sample appeared after 1500 msec fixation duration. The subject responded during the appearance of comparisons, 2700 msec, following the sample appearance of 800 msec. "O" indicated that the subject's response was correct, and "X" indicated wrong. (C) Testing order and time course of a session (block design). Baselines were AB in BA test, BC in CB test, and AB and BC in both AC and CA tests.

them were unknown to the subjects at the start of task. The subjects were instructed to learn the relations based on the auditory feedback they received after each keypress. During the second training session, the subjects were re-exposed to the same stimulus relations using the same procedure. The second training session always occurred prior to, and on the same day as, the test session.

During both training sessions, the subjects were seated 1 m in front of a monitor display. Each stimulus figure was 4 cm on each side on the display. The visual angle of the sample stimulus figure was 1.2°. The distance between the centers of the sample and comparison figures was 3.5 cm, and the visual angle to the outer edge of each comparison was 5.4° on the display. The subjects were instructed to indicate their selections on a Japanese keyboard (TK-UP87MPBK; ELECOM, Osaka, JP) by pressing the "a" key for left and the "]" key for right.

The training sessions were divided into blocks of 50 trials each. The subjects could take a rest in each interblock interval, and were able to initiate the next block by pressing any key. Blocks were administered repeatedly until each subject reached a criterion threshold of 90% accuracy (45 or more correct responses) in two consecutive blocks. That is, even if a subject performed the task perfectly in one block and then answered correctly in 44 trials in the subsequent block—thus achieving 94% accuracy across both trials—he or she was still deemed to have inadequately learned the relations. All subjects successfully reached this criterion in the first training session and returned on a subsequent day for the second training session and test session. The second training session used the same accuracy criterion as the first.

The trial format of the training sessions is called identity matching, which is widely used to confirm that subjects identify two instances of a stimulus that appear at different positions. The training blocks were designed to ensure that every stimulus figure appeared in each of the three possible positions (as sample, as left comparison, and as right comparison) during training. This was an essential control, because in order to probe transitivity and equivalence in the test session, the stimuli of Set A had to appear as comparisons and the stimuli of Set C had to appear as samples. If the subjects had not previously seen each figure in each position, we could not have ruled out the possibility that some of the associated activation was due to a figure appearing in a novel position, rather than due to the cognitive process of inference.

fMRI Experiment

Upon completing the second behavioral experiment, subjects were placed in an MRI scanner and were tested for their responses to the three emergent (and never-before-seen) relations of symmetry (BA and CB, illogical), transitivity (AC, logical), and their combination, equivalence (CA, illogical). The fMRI experiments used a different block design than behavioral experiments. Behavioral experiment blocks consisted of random mixtures of AB and BC trials, and the total number of blocks to which each subject was exposed in the behavioral experiments depended on his or her performance. The fMRI experiment, in contrast, contained a fixed number of blocks and was more carefully structured. The fMRI experiment was divided into four "sessions," each of which tested one of the four pairings representing the three types of emergent relation (Figure 2C). Each session contained five baseline blocks and five test blocks, presented in alternation. Each block contained 10 trials. Baseline blocks contained 10 baseline trials or learned relations (AB and BC). Test blocks contained five baseline trials and five test trials, or emergent relations. The BA session used BA for test trials and AB for baseline trials; the CB session used CB for test trials and BC for baseline trials; the AC session used AC for test trials and AB and BC for baseline trials; and the CA session used CA for test trials, and AB and BC for baseline trials. The first trial of each test block was always a test trial, to evoke the hemodynamic response associated with that relation, but the remaining four test trials and the five baseline trials were randomly ordered in the rest of the block. The reason we intermingled test trials and baseline trials within the test blocks was because we thought that blocks consisting entirely of test trials would have been too conspicuous. If the subjects had realized they were in a block of all-novel relations, they may have begun to behave or introspect differently.

In all trials of the test session, the subjects continued to choose between the presented comparison figures via button presses and continued to receive feedback. However, the feedback delivered in test trials was sham feedback: Regardless of whether their responses were correct, the subjects were invariably presented with a green "O" indicating "correct." This was done for two reasons. First, it prevented them from being able to learn the emergent relations by trial and error. Without this control, the inference-related brain activation of interest could have been contaminated by learning-related activation. Second, it provided both perceptual and psychological consistency between test trials and baseline trials. The subjects' implicit judgments about their selections were not controllable, but we knew they were already highly competent in baseline trials and had grown accustomed to believing that their responses were accurate most of the time. If we had not provided sham feedback during test trials, we might have invoked cognitive processes associated with a heightened level of reflection or self-doubt. By providing it, we hoped to keep the subjects' level of confidence in their responses comparable to what it was in baseline trials.

The four sessions were unvaryingly presented to all subjects in the sequence shown in Figure 2C. It was important to present the BA session before the CB session in order to probe for the cognitive process of "transfer." In this study, this refers to the possibility that the "rule" that the brain "discovered" during the BA session is applied (transferred)

to the CB session, with no need to "rediscover" it. Thus, if the transfer process was an essential component of symmetric inference in the CB session, we should observe a significant difference in both reaction time and response accuracy between the baseline trials and test trials in the BA session, but not in the CB session. More directly, we could check for behavioral data differences between the fifth test block in the BA session and the first block of the CB session. If such differences were observed, any brain areas that were activated in the BA session but not in the CB session would be implicated in transfer, which in turn would be informative about the cognitive process of symmetry.

In the scanner, the subjects responded by pressing buttons that were gripped separately in the left and right hands. The stimuli were presented through a fiber-optic goggle system (Avotec, Jensen Beach, FL) that subtended 30° × 23° of visual angle. The stimuli appeared near the center of vision so that the subjects could see them without eye movement. The visual angles subtended by the stimuli were about the same as they had been on the monitor display during the training sessions.

Image Acquisition

Images were collected using a 4-T Varian Unity Inova MRI system (Varian NMR Instruments, Palo Alto, CA). The BOLD signal was measured using a T2*-weighted echo-planar sequence (TR = 2600 msec, TE = 25 msec, FA = 40°). Twenty-five axial slices (thickness = 5.0 mm, FOV = 24 cm, 64 × 64 matrix, 0 mm gap) were acquired per volume. A high-resolution 3-D FLASH T1-weighted structural image was obtained (TR = 110 msec, TE = 6.2 msec, FA = 11°, 256 × 256 × 180 matrix, voxel size = 1 × 1 × 1 mm). Head motion was monitored, and respiration and cardiac signals were measured simultaneously, which were used in postprocessing for removal of physiological fluctuations (Hu, Le, Parrish, & Erhard, 1995) from functional images.

Imaging Data Analysis

Images were preprocessed and analyzed using Brain Voyager QX 1.8 software (Brain Innovation B.V., Maastricht, Netherlands). Functional images were spatially realigned in 3-D, spatially smoothed with a Gaussian filter (FWHM = 6 mm), stripped of linear trends, and temporally smoothed with a high-pass filter (3 cycles/points, about 0.0055 Hz). Slice-scan time correction was not performed because our block design did not require it. The structural image of each subject's brain was transformed into standard Talairach space. The cerebrum was translated and rotated into the AC–PC plane (AC = anterior commissure; PC = posterior commissure). The borders of the cerebrum were identified, and then the brain was resized (to 1 × 1 × 1 mm voxel size) and fitted into standard Talairach space. Functional images were transformed into standard Talairach space by normalization to the transformed structural image. BOLD signals were modeled using a synthetic hemodynamic response function composed of two gamma functions. The general linear model included two regressors, test and baseline. Contrast images (beta maps) were calculated and normalized to standard Talairach space. The test > baseline contrast shows the brain activation of emergent relation and cancels out that of baseline relation, because both test and baseline blocks contain baseline trials. Activations that survived cluster-level significance at a threshold of $p < .005$ (FDR corrected) and an extent threshold of 50 voxels were reported.

RESULTS

Behavioral Performance

Symmetry (BA and CB) and equivalence (CA) are logically invalid, whereas transitivity (AC) is valid. In accord with the extensive cognitive–psychological literature on implicit biases in human reasoning, the subjects systematically made illogical inferences in the symmetry and equivalence tests. For the purposes of analysis and discussion, these logically invalid responses are regarded as "accurate" or "correct." Response accuracy in each test session was significantly above chance (Figure 3A). Further analysis of response accuracy showed statistical significance. The main effect of baseline/test blocks was significant [$F(1, 14) = 42.5, p < .001$]. The main effect of test sessions was significant [$F(3, 42) = 14.9, p < .001$]. The interaction between baseline/test blocks and test sessions was significant [$F(3, 42) = 10.1, p < .001$]. Multiple comparisons of the main effect of test sessions showed four significant differences: BA and AC, BA and CA, CB and AC, CB and CA, namely, between the first two test sessions and the last two sessions. Multiple comparisons of the interaction revealed significant differences between test and baseline blocks in the BA, AC, and CA sessions, but not in the CB session.

Statistical analysis of reaction time showed that the main effect of baseline/test blocks was significant [$F(1, 14) = 83.9, p < .001$], the main effect of test sessions was significant [$F(3, 42) = 15.5, p < .001$], and their interaction was significant [$F(3, 42) = 8.13, p < .001$] (Figure 3B). Multiple comparisons of the main effect of test sessions showed results similar to those found in response accuracy. Multiple comparisons of the interaction also revealed similar results to those observed in response accuracy. Behavioral data of reaction time and response accuracy showed a significant difference between the baseline trials and test trials in the BA, AC, and CA sessions, but not in the CB session, as we mentioned that such difference among test sessions should be observed if the cognitive component of symmetric inference in the CB session was transferred from the BA session.

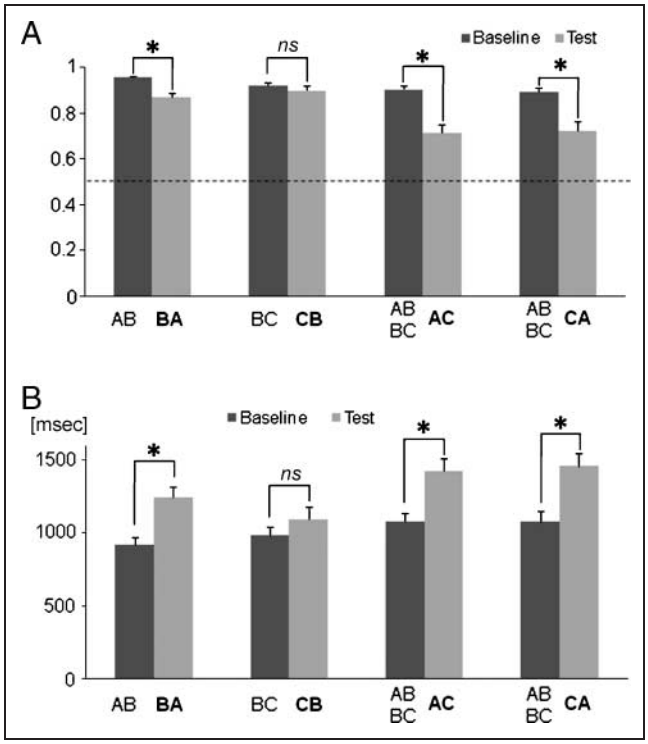

Figure 3. (A) Accuracy in each session. Chance level is shown by the dotted line. The correct rate in test trials was significantly above that in baseline trials except in the CB test session. The ranges show error bars indicate standard errors, $*p < .05$. (B) Reaction time in each session. The ranges and * indicate the same in the accuracy graph.

Further analysis of reaction time, a 2 × 2 ANOVA of baseline/test and session change (the fifth block of the BA session and the first block of the CB session), showed that the main effect of baseline/test and the interaction were significant [$F(1, 14) = 11.07$, $p < .01$, and $F(1, 14) = 4.97$, $p < .05$, respectively]. Multiple comparison of the interaction (Tukey's HSD test) showed that reaction time in the fifth test block of the BA session was not significantly different from that in the first test block of the CB session ($p > .05$). These results further strengthen the case for a transfer of symmetry from BA to CB.

Patterns of Brain Activation

Figure 4 shows brain activation during the symmetry sessions. Orange areas depict activation from the BA session (vs. baseline AB), and blue areas depict activation from the CB session (vs. baseline BC). In the BA session, bilateral prefrontal cortex (PFC), intraparietal area (IPA), and medial frontal cortex (MFC) were activated. In the CB session, only right PFC was activated. Brain activations in the BA and CB sessions that might reflect the process for symmetrical inference were different, as we expected that the difference of activation between the BA and CB session was observed. Figure 5 shows the activation from the transitivity and equivalence sessions. Orange areas depict activation from the AC session (vs. baselines AB, BC), and blue areas depict activation from the CA session (vs. baselines AB, BC). Both sessions showed brain activation similar to that in the BA session plus some additional activation: left dorsolateral PFC and right anterior PFC were activated in the AC session, and posterior cingulate cortex (PCC) and precuneus were activated in the CA session. Activated areas in each session are summarized in Table 1.

DISCUSSION

The subjects' response accuracy for each emergent relation was well above chance, demonstrating that they were "correctly" (i.e., as predicted, not with respect to logical validity) inferring symmetry, transitivity, and equivalence. Therefore, any contrast between the brain activation associated with test trials and baseline trials must be considered to reflect the processing of the emergent relations.

The subjects were apparently formulating their responses not based on the originally learned, unidirectional

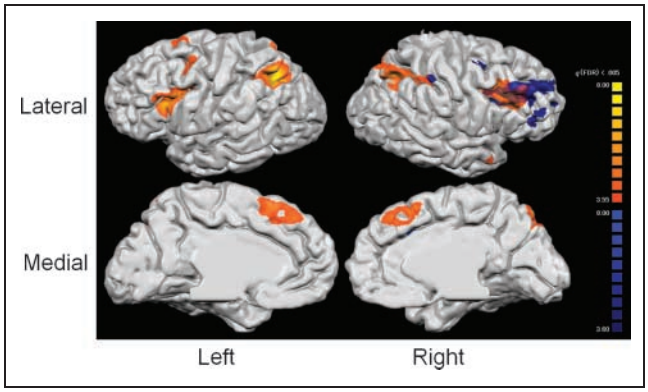

Figure 4. Activations associated with symmetry. The orange area shows the significant activation in the BA test session. Blue shows the activation in the CB session. Color bars show t values.

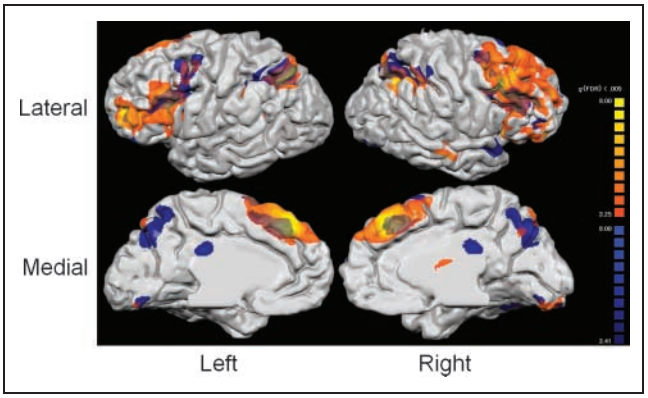

Figure 5. Activations associated with transitivity and equivalence. The orange area shows the significant activation in the AC test session. Blue shows the significant activation in the CA test session. Color bars show t values.

Table 1. Contrast between Activations (Test > Baseline) with a Statistical Threshold of $p < .005$ (FDR Corrected) and an Extent Threshold of 50 Voxels

	BA		*CB*		*AC*		*CA*	
Location	*Talairach Coordinates*	*t*	*Talairach Coordinates*	*t*	*Talairach Coordinates*	*t*	*Talairach Coordinates*	*t*
L IPA	−33, −61, 43	8.12			−33, −64, 43	8.56	−39, −52, 49	7.71
R IPA	30, −52, 43	7.13			30, −67, 43	7.55	26, −64, 40	7.13
MFC	−6, 14, 49	5.57			0, 32, 43	9.98	0, 26, 43	6.80
L PFC	−49, 20, 31	6.20			−45, 20, 31	8.76	−45, 17, 34	6.64
R PFC	42, 28, 28	7.26	33, 38, 28	6.47	39, 17, 40	10.90	36, 35, 22	8.06
L aPFC					30, 47, 7	8.99		
PCC							0, −28, 25	5.61
Precuneus					−6, −64, 40	4.78	−3, −55, 52	5.46

L = left; R = right; a = anterior; PCC = posterior cingulate cortex.

stimulus relations, but on bidirectional categories derived from them. (The trained relations were unidirectional insofar as AB and BC were defined as correct, while BA and CB were undefined, rather than BA and CB being defined as incorrect. The point was not to extinguish the symmetrical relations, but to demonstrate that humans automatically infer them, in contrast to nearly all other animals that have been studied under comparable paradigms.) In the framework of stimulus equivalence, these abstract categories are called equivalence classes (Kastak, Schusterman, & Kastak, 2001; Sidman & Tailby, 1982). Activated brain regions and their associated cognitive functions were assessed with this framework in mind.

Fronto-parietal Network

Our data lent strong support to the idea that the cognitive function for symmetry was recruited in the BA session and transferred to the CB session. Firstly, the behavioral data show that reaction time and response accuracy did not differ significantly between the fifth BA block and the first CB block. This observation accords with other studies on cognitive transfer. Adult and child subjects who were exposed to conditional discrimination in a matching-to-sample task showed accurate responses in the sense of "illogically inferred, just as predicted," when tested for the first time for transfer (Perez-Gonzalez, 1994; Green, Sigurdardottir, & Saunders, 1991).

Secondly, bilateral IPA activation was observed in all sessions except CB. IPA has been suggested to encode and mediate the grouping of visual stimuli. One study showed that grouped stimuli invoke lower activation in IPA than ungrouped stimuli (Xu & Chun, 2007), which might have the effect of reducing the cognitive load required to process grouped stimuli. Superior IPA might be associated with the number of objects stored in visual short-term memory (Xu & Chun, 2006; Todd & Marois, 2004). A neurophysiological study of monkeys showed that IPA activation encoded category information about motion direction of dots, and that the activation shifted to encode the newly trained category after retraining the monkeys to group the same motion directions into different categories (Freedman & Assad, 2006). Thus, we suggest that the bilateral IPA activation in our study reflects the active, immediate manipulation of the learned relations to make inferences about the novel relations (see below for defense of this interpretation). Specifically, we suggest this manipulation brings about the inversion or reversal of the learned relations to derive symmetry. This would explain why IPA activation was not observed in the CB session (by dint of transfer).

In contrast to IPA, right PFC activation was observed in both the BA and CB sessions. Right PFC is sensitive to category learning (Reber, Stark, & Squire, 1998) and changes of category membership (Jiang et al., 2007), suggesting that the right PFC activation in our study may reflect the manipulation of relations among members of a category—in this case, an equivalence class. In test trials, the well-learned, temporo-spatially unidirectional relations of AB and BC were bidirectionalized by symmetrical inference in the BA and CB sessions, respectively. Thus, the observed PFC activation might involve the re-representation or modulation the visual-grouping information generated in parietal cortex in accordance with the demands of our experimental task.

PFC activation has also been associated with temporal context memory and decision-making based on it (Suzuki et al., 2002), and it has been suggested that parietal cortex mediates the representation of temporal order information (Marshuetz, Reuter-Lorenz, Smith, Jonides, & Noll, 2006). The prefrontal and parietal cortices might be cooperatively involved in the processing of working memory for temporal order information (Marshuetz & Smith, 2006; Marshuetz, Smith, Jonides, DeGutis, & Chenevert,

2000). Marshuetz et al. (2006) also observed that left parietal activation correlated to spatial inter-item distance. In our experiment, the stimulus-presentation protocol imposed a temporo-spatial order structure on the stimulus relations our subjects were asked to learn and, subsequently, infer. Thus, it may not be surprising that this fronto-parietal network was activated by our task.

Medial Frontal Cortex

MFC was activated differentially by exposure to learned and emergent relations, indicating that the subjects were detecting the difference between these two types of relation. By the time of the test session, the subjects were familiar with all the stimulus figures as well as the defined relations between them. Also, by design, they had been exposed to each figure appearing in each of the three possible stimulus positions (left, center, right) on the display in the behavioral experiments. However, they had not been exposed to every possible combination of three figures. By necessity, the unlearned relations in test trials consisted of never-before-seen combinations. Neuroimaging studies using task switching and stimulus–response conflict paradigms have demonstrated that MFC is sensitive to conflict and error (Yeung & Cohen, 2006; Rushworth, Hadland, Paus, & Sipila, 2002). A meta-analysis of MFC function suggests that MFC is associated with the detection of unfavorable outcomes, response errors, response conflict, and decision uncertainty (Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). MFC and anterior cingulate cortex, which lies inferior to MFC, have been associated with the monitoring of response conflict and error detection (Ridderinkhof et al., 2004). We suggest that the MFC activation in our study reflects additional processing for emergent relations relative to baseline triggered by the conditional novelty of the emergent-relation tests.

Many studies have implicated MFC in error-related processing (Bengtsson, Lau, & Passingham, 2009; Hester, Barre, Murphy, Silk, & Mattingley, 2008; Krigolson & Holroyd, 2007; Nieuwenhuis, Schweizer, Mars, Botvinick, & Hajcak, 2007; Mars et al., 2005; Ridderinkhof et al., 2004; van Schie, Mars, Coles, & Bekkering, 2004; Gehring & Willoughby, 2002). A possible concern with the block design of our experiment is that it did not include controls that would allow us to explicitly exclude or isolate possible error-related activation. The number of feedbacks was the same in test blocks and baseline blocks in a session, indicating that the feedback-related brain activation would be canceled out in the test > baseline contrast. Analyzing the behavioral data, the average number of positive feedback signals delivered throughout all five baseline blocks and all five test blocks were 46.02 (of 50) and 47.97 (of 50), respectively. (This includes the all-positive sham feedback delivered in test trials, which comprised half the trials within each test block.) A binominal test was performed to compare the number of positive feedback signals in test blocks with that in baseline blocks as the ideal number. The difference was not significant ($p > .05$). This suggests, although it does not prove, that the observed MFC activation is not due to error processing related to feedback.

Finally, it has been suggested that pre-SMA, which was activated in all sessions except CB, is associated with switching behavioral rules (Isoda & Hikosaka, 2007; Rushworth et al., 2002). Significant differences in response accuracy and reaction time were observed between baseline trials and test trials in BA, AC, and CA sessions, suggesting some cost for emergent relations. However, as was expected, no significant differences in response accuracy or reaction time were observed in the CB session. This suggests that MFC activation might be associated with a cognitive processing activated by the conditional novelty of emergent relation after all.

Medial Parietal Cortices

In order to formulate correct responses in the AC and CA sessions, the subjects had to retrieve information about the unseen intermediary stimulus, B. The precuneus was activated in the AC session, whereas both the precuneus and PCC were activated in the CA session. Activation of PCC and precuneus has been associated with the process of recalling memorized stimuli (Schott et al., 2005; Acuna, Eliassen, Donoghue, & Sanes, 2002; Henson, Rugg, Shallice, Josephs, & Dolan, 1999). Findings of the activation in medial temporal cortex, including hippocampus in transitivity and equivalence, suggested that the area was involved in maintaining relational structure and flexible memory expression in an equivalence class (Schlund et al., 2008). It is plausible that the activation was related to short-term memory associated with hippocampus, because the subjects might easily maintain six phonic stimuli, consonant, and vowel, in their short-term memory during fMRI scan and the period of 3 hours between training and fMRI scan. In our study, there was at least one night between training and fMRI scan, suggesting that the subjects memorized the stimulus relations into their long-term memory. Therefore, we suggest that these activation patterns are associated with the recall of B.

Inference but not Learning

A key assumption in our interpretation of our results is that the subjects were immediately inferring each novel relation in each test trial. The difference in the behavioral data between baseline and test, presented earlier, supports this. However, one possible explanation of our results is that the subjects did not immediately infer, but rather learned, the emergent relations. In each test block, the subjects were exposed to all five distinct test relations. In line with this possibility, the subject that did not infer should choose randomly in the first exposure of test trials.

Conceivably, the sham feedback delivered in test trials could have led to immediate learning of each relation upon first exposure in the first test block. In the remaining four test blocks, the subjects could have been applying these newly learned behavioral patterns to subsequent instances of each relation. If immediate learning was the explanation, its effect should have been observed as equivalent accuracy in the first and the other test trials of each relation in the first and the other test blocks. However, response–accuracy analysis showed significant main effect of test blocks [ANOVA of test sessions and test blocks, $F(4, 56) = 7.14$, $p < .01$], suggesting a fluctuation of accuracy among test blocks. Because the sham feedback was positive irrespective of response accuracy, learning through trial and error was not possible. Therefore, the best explanation of our data is that the subjects were performing symmetrical and transitive inference to make their selection in each test trial.

As described above, the cognitive process for symmetry might transfer directly from the BA session to the CB session. It is also conceivable, however, that the subjects might have switched rules or pattern of behavior sometime during the BA session itself. To test for this, we analyzed the behavioral data to see if the ratio of reaction time of baseline and test (baseline/test) changed significantly. (We excluded the initial period of adaptation to the task in the MRI scanner, which occurred in both of baseline and test.) One-way ANOVA of blocks in the BA session showed no significant difference [$F(4, 56) = 0.274$, $p = .89$], suggesting that the subjects were stably applying symmetry throughout the BA session. Another possible concern was that BA was automatically inferred during the training of AB; in other words, the subjects had "learned" both AB and BA before entering the scanner, where they were exposed to BA explicitly for the first time. This concern was eliminated by the fact that the behavioral data of AB and BA were significantly different, suggesting that BA was inferred in the test but not in the training.

Similarity of Brain Activation between Logical and Illogical Inferences

One of main objectives in this study was to discern the neural commonalities and differences between human logical and illogical inference. It was apparently supposed that illogical inference that meant not necessarily true was associated with brain activation that was not shared with the activation of logical inference. Activations in PFC and IPA were observed during the cognitive inferential processes of symmetry, transitivity, and equivalence. Additional activation in the precuneus and PCC during transitivity and equivalence might correspond to memory retrieval processes. Surprisingly, aside from some activation in left anterior PFC during transitive inference and PCC during equivalence inference, two patterns of brain activation in illogical and logical inferences were very similar. Therefore, we suggest that PFC and IPA subsume a uniquely human fronto-parietal network that is centrally involved in categorizing spatially, temporally, and causally related visual stimuli, and in organizing them to form inferences. The inferences formed by this network are based on associative category membership rather than on any unidirectional structure in the original relations between the relevant events or stimuli. Thus, whether these inferences happen to be logically valid or fallacious in the strict, syllogistic sense is purely incidental. The human-typical pattern of logical and illogical inference falls out as a natural consequence of this interpretation. The adaptive advantage conferred by this innate "naive" style of reasoning was perhaps to reduce the cognitive load on working memory (Burgess, 1999), and thus, allow humans to manage and process much greater volumes of information.

And yet, people with training and education are able to reason formally without falling prey to these "natural" cognitive biases. What accounts for this? Several studies have looked at brain activation during logical reasoning. Before receiving training in logical reasoning, subjects showed activation in posterior parts of the brain (occipital and parietal cortices); but after such training, they showed frontal activation, which was interpreted as suppressing the naturally biased responses that were recognized as formally incorrect (Houdé, 2007; Houdé & Tzourio-Mazoyer, 2003; Houdé et al., 2000). A PET study investigated the effect of emotion and feeling on deductive logic. A subject group that underwent logico-emotional training showed an error-to-logical shift, which was not observed in the other subject group that was trained in logic only. The between-group comparison showed differential activation in ventromedial PFC, suggesting that this emotion-related area facilitates deductive reasoning by suppressing the naturally biased response (Houdé & Tzourio-Mazoyer, 2003; Houdé et al., 2001). Another study found that logical reasoning specifically recruited mid-dorsolateral PFC, MFC, and parietal cortices when perceptual features biased subjects' responses toward being incorrect (Prado & Noveck, 2007). These results suggest that PFC and MFC support formal logical reasoning by suppressing the natural cognitive biases that are associated with parietal and occipital regions. These studies accord well with our results presented here, which suggest that a fronto-parietal network is associated with logical inference, transitivity, and biased inferences, symmetry and equivalence.

This experiment did not address the question of what combination of genetically and culturally organized brain modifications enable humans to perform rigorous, syllogistically sound inference. Nor did it directly address the relationship or overlap between the human performance of unimodal *visual* inference and propositional *linguistic* inference. However, it may be noteworthy that the species-typical biases of symmetry, transitivity, and equivalence manifest similarly when humans perform inference in both of these domains.

Reprint requests should be sent to Atsushi Iriki, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan, or via e-mail: iriki@brain.riken.jp.

REFERENCES

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

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,* 204–210.

Bengtsson, S. L., Lau, H. C., & Passingham, R. E. (2009). Motivation to do well enhances responses to errors and self-monitoring. *Cerebral Cortex, 19,* 797–804.

Boysen, S. T., Berntson, G. G., Shreyer, T. A., & Quigley, K. S. (1993). Processing of ordinality and transitivity by chimpanzees (Pan troglodytes). *Journal of Comparative Psychology, 107,* 208–215.

Burgess, N. G. J. H. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106,* 551–581.

D'Amato, M. R., Salmon, D. P., Loukas, E., & Tomie, A. (1985). Symmetry and transitivity of conditional relations in monkeys (*Cebus apella*) and pigeons (*Columba livia*). *Journal of the Experimental Analysis of Behavior, 44,* 35–47.

Dickins, D. W. (2005). On aims and methods in the neuroimaging of derived relations. *Journal of the Experimental Analysis of Behavior, 84,* 453–483.

Dickins, D. W., Singh, K. D., Roberts, N., Burns, P., Downes, J. J., Jimmieson, P., et al. (2001). An fMRI study of stimulus equivalence. *NeuroReport, 12,* 405–411.

Dickins, E. T., & Dickins, D. W. (2001). Symbols, stimulus equivalence and the origin of language. *Behavior and Philosophy, 29,* 221–244.

Eikeseth, S., & Smith, T. (1992). The development of functional and equivalence classes in high-functioning autistic children: The role of naming. *Journal of the Experimental Analysis of Behavior, 58,* 123–133.

Freedman, D. J., & Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature, 443,* 85–88.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295,* 2279–2282.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences, 11,* 435–441.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage, 12,* 504–514.

Green, G., Sigurdardottir, Z. G., & Saunders, R. R. (1991). The role of instructions in the transfer of ordinal functions through equivalence classes. *Journal of the Experimental Analysis of Behavior, 55,* 287–304.

Hank, D. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology, 106,* 342–349.

Henson, R. N., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience, 19,* 3962–3972.

Hester, R., Barre, N., Murphy, K., Silk, T. J., & Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cerebral Cortex, 18,* 1933–1940.

Houdé, O. (2007). First insights on "neuropedagogy of reasoning". *Thinking & Reasoning, 13,* 81–89.

Houdé, O., & Tzourio-Mazoyer, N. (2003). Neural foundations of logical and mathematical cognition. *Nature Reviews Neuroscience, 4,* 507–514.

Houdé, O., Zago, L., Crivello, F., Moutier, S., Pineau, A., Mazoyer, B., et al. (2001). Access to deductive logic depends on a right ventromedial prefrontal area devoted to emotion and feeling: Evidence from a training paradigm. *Neuroimage, 14,* 1486–1492.

Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., et al. (2000). Shifting from the perceptual brain to the logical brain: The neural impact of cognitive inhibition training. *Journal of Cognitive Neuroscience, 12,* 721–728.

Hu, X., Le, T. H., Parrish, T., & Erhard, P. (1995). Retrospective estimation and correction of physiological fluctuation in functional MRI. *Magnetic Resonance in Medicine, 34,* 201–212.

Isoda, M., & Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience, 10,* 240–248.

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., & Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron, 53,* 891–903.

Kastak, C. R., & Schusterman, R. J. (2002a). Long-term memory for concepts in a California sea lion (*Zalophus californianus*). *Animal Cognition, 5,* 225–232.

Kastak, C. R., & Schusterman, R. J. (2002b). Sea lions and equivalence: Expanding classes by exclusion. *Journal of the Experimental Analysis of Behavior, 78,* 449–465.

Kastak, C. R., Schusterman, R. J., & Kastak, D. (2001). Equivalence classification by California sea lions using class-specific reinforcers. *Journal of the Experimental Analysis of Behavior, 76,* 131–158.

Krigolson, O. E., & Holroyd, C. B. (2007). Predictive information and error processing: The role of medial-frontal cortex during motor control. *Psychophysiology, 44,* 586–595.

Mars, R. B., Coles, M. G., Grol, M. J., Holroyd, C. B., Nieuwenhuis, S., Hulstijn, W., et al. (2005). Neural dynamics of error processing in medial frontal cortex. *Neuroimage, 28,* 1007–1013.

Marshuetz, C., Reuter-Lorenz, P. A., Smith, E. E., Jonides, J., & Noll, D. C. (2006). Working memory for order and the parietal cortex: An event-related functional magnetic resonance imaging study. *Neuroscience, 139,* 311–316.

Marshuetz, C., & Smith, E. E. (2006). Working memory for order information: Multiple cognitive and neural mechanisms. *Neuroscience, 139,* 195–200.

Marshuetz, C., Smith, E. E., Jonides, J., DeGutis, J., & Chenevert, T. L. (2000). Order information in working memory: fMRI evidence for parietal and prefrontal mechanisms. *Journal of Cognitive Neuroscience, 12,* 130S–144S.

Nieuwenhuis, S., Schweizer, T. S., Mars, R. B., Botvinick, M. M., & Hajcak, G. (2007). Error-likelihood prediction in the medial frontal cortex: A critical evaluation. *Cerebral Cortex, 17,* 1570–1581.

O'Donnell, J., & Saunders, K. J. (2003). Equivalence relations in individuals with language limitations and mental retardation. *Journal of the Experimental Analysis of Behavior, 80,* 131–157.

Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex, 11,* 954–965.

Perez-Gonzalez, L. A. (1994). Transfer of relational stimulus control in conditional discriminations. *Journal of the Experimental Analysis of Behavior, 61,* 487–503.

Prado, J., & Noveck, I. (2007). Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience, 19,* 642–657.
Reber, P. J., Stark, C. E., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences, U.S.A., 95,* 747–750.
Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306,* 443–447.
Rushworth, M. F., Hadland, K. A., Paus, T., & Sipila, P. K. (2002). Role of the human medial frontal cortex in task switching: A combined fMRI and TMS study. *Journal of Neurophysiology, 87,* 2577–2592.
Saunders, R. R., & Green, G. (1999). A discrimination analysis of training-structure effects on stimulus equivalence outcomes. *Journal of the Experimental Analysis of Behavior, 72,* 117–137.
Schlund, M. W., Cataldo, M. F., & Hoehn-Saric, R. (2008). Neural correlates of derived relational responding on tests of stimulus equivalence. *Behavioral and Brain Functions, 4,* 6.
Schlund, M. W., Hoehn-Saric, R., & Cataldo, M. F. (2007). New knowledge derived from learned knowledge: Functional–anatomic correlates of stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 87,* 287–307.
Schott, B. H., Henson, R. N., Richardson-Klavehn, A., Becker, C., Thoma, V., Heinze, H. J., et al. (2005). Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering, and control of retrieval. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 1257–1262.
Schusterman, R. J., & Kastak, D. (1998). Functional equivalence in a California sea lion: Relevance to animal social and communicative interactions. *Animal Behavior, 55,* 1087–1095.
Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37,* 5–22.
Suzuki, M., Fujii, T., Tsukiura, T., Okuda, J., Umetsu, A., Nagasaka, T., et al. (2002). Neural basis of temporal context memory: A functional MRI study. *Neuroimage, 17,* 1790–1796.
Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428,* 751–754.
Tomonaga, M., Matsuzawa, T., Fujita, K., & Yamamoto, J. (1991). Emergence of symmetry in a visual conditional discrimination by chimpanzees (*Pan troglodytes*). *Psychological Reports, 68,* 51–60.
Treichler, F. R., & Van Tilburg, D. (1996). Concurrent conditional discrimination tests of transitive inference by macaque monkeys: List linking. *Journal of Experimental Psychology: Animal Behavior Processes, 22,* 105–117.
van Schie, H. T., Mars, R. B., Coles, M. G., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience, 7,* 549–554.
von Fersen, L., Wynne, C. D. L., Delius, J. D., & Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 17,* 334–341.
Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440,* 91–95.
Xu, Y. D., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences, U.S.A., 104,* 18766–18771.
Yamazaki, Y. (2004). Logical and illogical behavior in animals. *Japanese Psychological Research, 46,* 195–206.
Yeung, N., & Cohen, J. D. (2006). The impact of cognitive deficits on conflict monitoring. Predictable dissociations between the error-related negativity and N2. *Psychological Science, 17,* 164–171.