LATERALITY, 2006, 11 (3), 195–225

Psychology Press
Taylor & Francis Group

Theory of mind and the right cerebral hemisphere: Refining the scope of impairment

Richard Griffin
Tufts University, Boston, MA, USA
Ori Friedman
University of Waterloo, Ontario, Canada
Jon Ween
University of Toronto Faculty of Medicine, Canada
Ellen Winner
Boston College, MA, USA
Francesca Happé
King's College London, UK
Hiram Brownell
Boston College, MA, USA

The neuropsychological and functional characterisation of mental state attribution (''theory of mind'' (ToM)) has been the focus of several recent studies. The literature contains opposing views on the functional specificity of ToM and on the neuroanatomical structures most relevant to ToM. Studies with brain-lesioned patients have consistently found ToM deficits associated with unilateral right hemisphere damage (RHD). Also, functional imaging performed with non-brain-injured adults implicates several specific neural regions, many of which are located in the right hemisphere. The present study examined the separation of ToM impairment from other deficits associated with brain injury. We tested 11 patients with unilateral right hemisphere damage (RHD) and 20 normal controls (NC) on a

Address correspondence to: Richard Griffin, Center for Cognitive Studies, Tufts University, 11 Miner Hall, Medford, MA 02155, USA. E-mail: richard.griffin@tufts.edu

This research and writing was supported by grants R01 NS 27894, P30 DC 05207, R01 DC 05432, and an MRC program grant (to Simon Baron-Cohen and Mark Johnson). The first author was further supported by a fellowship from the Cambridge Overseas Trust. Thanks to Ana Raposo at the Centre for Speech and Language in Cambridge for help with the images. We would also like to thank all the participants and their families for their time and generosity.

http://www.psypress.com/laterality DOI: 10.1080/13576500500450552

humour rating task, an emotion rating task, a graded (first-order, second-order) ToM task with non-mentalistic control questions, and two ancillary measures: (1) Trails A and B, in order to assess overall level of impairment and set-shifting abilities associated with executive function, and (2) a homograph reading task to assess central coherence skills. Our findings indicate that RHD can result in a functionally specific deficit in attributing intentional states, particularly those involving second-order attributions. Performance on ToM questions was *not* reliably related to measures of cognitive impairment; however, performance on non-ToM control questions was reliably predicted by Trails A and B. We also discuss individual RHD patients' performance with attention to lesion locus. Our findings suggest that damage to the areas noted as specialised in neuroimaging studies may not affect ToM performance, and underscore the necessity of combining lesion and imaging studies in determining functional-anatomical relations.

The ability to attribute mental states ("theory of mind") to ourselves and others represents a fundamental and distinctive component of human social cognition (Dennett, 1996; Heyes, 1998; Povinelli, 2000; Tomasello, 1998). As we illustrate below, the representational capacities underlying "theory of mind" (ToM) figure in one's ability to deceive, to joke, and to engage in successful social interactions more generally. Complex cognitive abilities such as those included under the umbrella term "ToM" will of course call on several structures in the brain. In this paper, we present a neuropsychological investigation to examine more precisely the role of the right cerebral hemisphere and, more generally, to explore the parameters of disruption to ToM. By basing our investigation on non-aphasic patients with right hemisphere damage (RHD), we are able to examine ToM using a variety of performance measures.

The right hemisphere is a plausible location for certain components of ToM because there are striking parallels in symptomatology between those with high functioning autism/Asperger's syndrome, who are thought to lack or have deficient ToM, and patients with RHD. Both populations show deficits in the production and comprehension of non-literal speech (metaphor and irony), in the production and comprehension of humour, the perception and expression of emotion, the production and comprehension of prosody, and inferring and integrating verbal and pictorial information, and both groups show abnormal social behaviour (Brownell, Griffin, Winner, Friedman, & Happé, 2000; see also Happé, Brownell, & Winner, 1999, Table 1, and Sabbagh, 1999, for a catalogue of these shared impairments.)

There is also neuropsychological evidence consistent with the view that ToM relies, at least in part, on the RH. First, several studies have shown that RHD patients show specific deficits on ToM tasks, whereas patients with unilateral LH lesions do not generally show impairments. Most relevant to

the current study, Happé et al. (1999) directly compared individuals with unilateral LHD (with aphasia) and unilateral RHD on ToM and non-ToM inference tasks using single-frame cartoons and short stories. The RHD but not the LHD group showed selective difficulties on the ToM measures. Similarly, Rowe, Bullock, Polkey, and Morris (2001) tested unilateral RHD and LHD frontal patients on ToM and executive function tasks, and while they found that both groups had some difficulties with mental state attribution, the LHD frontal group also showed impairments in executive function tasks such as the Wisconsin Card Sort (WCST), the Stroop test, and Trailmaking (see also Channon & Crawford, 2000). Likewise, Stone, Baron-Cohen, and Knight (1998) tested a group of patients with unilateral left dorsolateral prefrontal damage and another group with bilateral orbitofrontal damage on a series of developmentally graded ToM tasks (first-order, second-order, and faux pas detection). First-order theory of mind (first-order ToM) tasks require an individual to determine a character's belief about the world, whereas passing a second-order theory of mind tasks (second-order ToM) requires the ability to determine the content of someone's belief about another person's belief. In the Stone et al. (1998) study, neither group had difficulty with the first-order or second-order items when memory loads were controlled for. The bilateral orbitofrontal group, however, had more difficulty on the faux pas detection task than did the LHD group. Finally, Stuss, Gallup, and Alexander (2001) found that right frontal (ventral) lesions impaired the detection of deception and that bilateral lesions impaired visual perspective taking (see also Shamay-Tsoory, Tomer, Berger, & Aharon-Peretz, 2003). Varley and Siegal (2000) tested a severely aphasic patient with left hemisphere brain damage (LHD) on first-order and second -order ToM tasks and found no deficit, despite an almost total absence of usable language. Taken together, these results suggest an important role for RH regions in components of reasoning that may be critical to ToM.

Additional evidence that RH structures are relevant to ToM is provided by findings of RH activation in several ToM imaging studies performed with non-brain-damaged participants, although activity is typically found in both hemispheres (Calarge, Andreasen, & O'Leary, 2003). Specific regions include the orbitofrontal cortex (Brodmann's area [BA] 10–14), cingulate gyri (BA 24, 31, 32), the right middle frontal gyrus (BA 6), the superior temporal gyrus (BA 22, 38, 39), the temporo-parietal junction, and the precuneus (BA 7) (Brunet, Sarfati, Hardy-Baylé, & Decety, 2000; Fletcher et al., 1995; Gallagher, Happé, Brunswick, Fletcher, Frith, & Frith, 2000; Goel, Grafman, Sadato, & Hallett, 1995; Happé et al., 1996; Saxe & Kanwisher, 2003). The majority of this imaging work has noted activation in and around the medial frontal cortex/paracingulate gyrus (near BA 8 and 9) (for a review see Gallagher & Frith, 2003). Interestingly, Bird, Castelli, Malik, Frith, and

Husain (2004) recently tested a patient with extensive bilateral damage to this area but found no ToM deficit, despite several executive function difficulties.

Most relevant to the current study are findings by Brunet et al. (2000) and Gallagher et al. (2000) who used single-frame wordless cartoons, including some of the same items used in this study. Brunet et al. found RH activation specific to the ToM items in the right medial frontal/paracingulate cortex (BA 6, 8, 9), anterior cingulate (BA 24), right middle frontal gyrus (BA 8/9), and right inferior temporal gyrus (BA 20/21). Gallagher et al. found significant activation specific to the ToM cartoons in the right middle frontal gyrus (BA 6), precuneus (BA 7/31), and the cerebellum. These areas are depicted in Figure 1. Moreover, right but not left orbitofrontal areas have been activated in a ToM neuroimaging (SPECT) study by Baron-Cohen, Ring, Moriarty, Schmitz, Costa, and Ell (1994). This region is considered to be part of a dedicated ToM circuit by several authors (for a review see Griffin & Baron-Cohen, 2002).

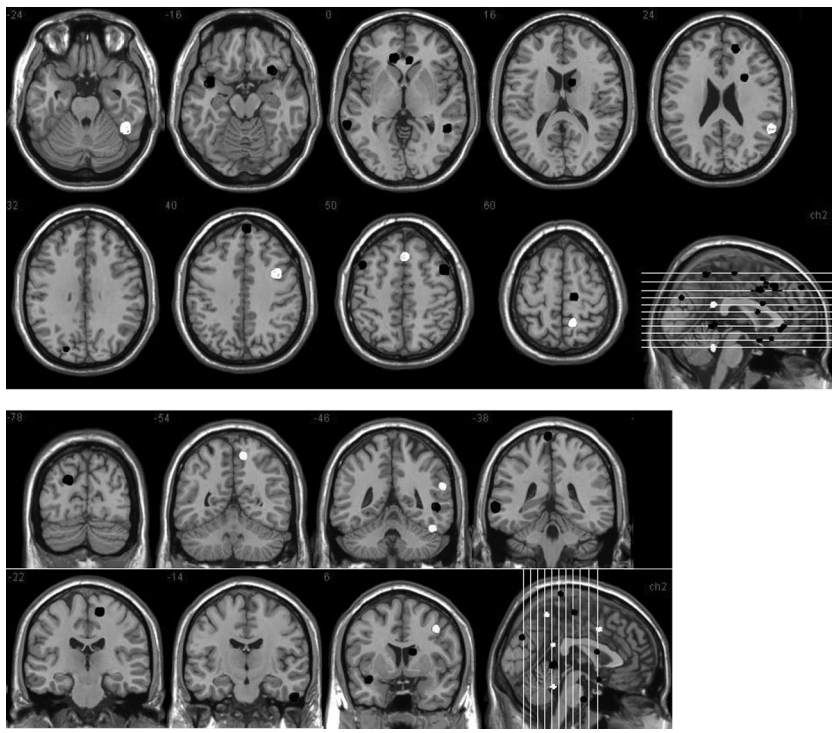

Figure 1. Selective ToM activation from two studies using nonverbal (cartoon) stimuli (White = Gallagher et al., 2000; Black = Brunet et al., 2000).

The ToM literature contains extensive discussion of the extent to which ToM can be distinguished from other domains of cognition. Claims range from versions of modularity with dedicated substrates (e.g., Baron-Cohen, 1995; Frith & Frith, 1999) to the view that ToM relies on domain-general executive functions (e.g., Ozonoff, Pennington, & Rogers, 1991; Pennington & Ozonoff, 1996; Pennington, Rogers, Bennetto, Griffith, Reed, & Shyu, 1997; Russell, 1997). Some authors take a middle ground, claiming that only some ToM deficits are best viewed as the result of general neurological and cognitive decline (e.g., Zaitchik, Koff, Brownell, Winner, & Albert, 2004). It is also possible, and indeed likely, that ToM reasoning could be domain specific at the cognitive level while relying on neurological substrates shared with other domains. This position is not logically inconsistent in that cognitive and neurological descriptions proceed at different levels.

Among theoretical accounts of ToM impairments in autism, two domain-general deficits have been proposed. "Weak central coherence" was proposed by Frith (1989), to explain both assets and impairments in autism. Central coherence refers to the ability to integrate information at different levels, e.g., extracting the gist or meaning of a story from the surface form. Weak coherence refers to a bias to attend to parts rather than wholes, reflected in piecemeal processing that is relatively context independent. This bias has been demonstrated, for example, in reading homographs (words with the same spelling but different meanings and pronunciations, such as "read" and "tear"). The meaning (and thus pronunciation) of a homograph is disambiguated by the context in which it appears, e.g., "In Sally's [eye/dress] there was a big tear." Individuals with autism have difficulty in using context to disambiguate homographs (Frith & Snowling, 1983; Happé, 1997; Jolliffe & Baron-Cohen, 1999). This detail-oriented processing style can be found in other domains as well, as Frith's use of "central" suggests. Individuals with autism show unusually good performance on visuospatial tasks such as the Embedded Figures Test and the Wechsler Block Design; both require attention to local rather than global information (Jolliffe & Baron-Cohen, 1997; Shah & Frith, 1983, 1993). The local versus global distinction has long been used to characterise the processing styles of the left and right hemispheres respectively (Springer & Deutsch, 1998). This is not only true with language; there is evidence that RHD patients' integrative processing deficits in the visuo-spatial domain correlate with integrative deficits in the verbal-semantic domain (Benowitz, Moya, & Levine, 1990). Frith suggested that weak coherence underlies ToM deficits in autism, and there is some evidence of a (negative) relation between the two constructs (Jarrold, Butler, Cottington, & Jimenez, 2000; but see discussion in Happé, 2000).

A second domain-general explanation of ToM deficits in autism has been executive dysfunction (for a review see Hill, 2004). Problems in inhibition and set shifting are prominent in (although not specific to) autism, and have

been proposed to underlie ToM task failure (see also Shallice, 2001). Pennington and Ozonoff's (1996) review of autistic performance of EF tasks found significant impairment in 25 of 32 EF tasks administered, compared to IQ-matched controls. Ozonoff et al. (1991) argue that performance on EF tasks is a better diagnostic of autism than performance on ToM tasks. Measures of coherence and of executive function (EF) were therefore included in the present study to examine their relation to predicted ToM deficits in RHD. While these two domain-general explanations do not exhaust the possible accounts, our aim was to test whether ToM could be distinguished from non-specific impairments broadly, which would lend further support to the claims that ToM is cognitively domain specific. Our secondary aim was to test whether damage to any of the brain regions showing significant activation during ToM imaging studies would affect performance, and in what ways, thereby refining the scope of ToM impairment following RHD.

In other respects, the study reported here builds directly on the study by Happé et al. (1999). Happé et al. tested groups of patients with unilateral RHD (without aphasia) and LHD (with Broca's aphasia) on single-frame cartoons to test the role for an intact right hemisphere in ToM. The RHD, but not the LHD, group had significantly more difficulty on the ToM items compared to the non-mental control items. Like Happé et al. (1999), we tested a group of RHD patients and a group of non-brain-damaged control participants using some of the original cartoons as well as additional cartoons. However, we modified Happé et al.'s study in several ways. First, we tested only RHD patients and non-brain-damaged control participants in order to allow maximum flexibility with respect to the linguistic demands of our task. This strategy allowed examination of ToM from a number of perspectives.

More substantively, we included an important distinction between two types or levels of ToM ability in our humour items: first-order belief (understanding a character's mental state) and second-order belief (understanding what one character believes about another character's beliefs). Second-order beliefs, which are fundamental to deception and sarcasm, are more complex and may be particularly difficult for RHD patients (Brownell et al., 2000; Stone et al., 1998). To gain perspective on the levels of ToM, we included three types of cartoon items: non-ToM items, first-order ToM items, and second-order ToM items. In addition, after *all* types of items we posed different comprehension questions to participants: we queried elements of knowledge/ignorance, which requires first-order processing, and deception, which requires second-order processing. As a control for the abstract nature of these mentalist notions, after each item we also asked about the danger and possibility/likelihood of events portrayed in the cartoon. Moreover, in contrast to Happé et al. (1999) and Rowe et al. (2001), we asked both open-ended and forced-choice questions to rule out the

possibility that the RHD patients had difficulties with articulating the relevant factors, rather than with ToM per se.

Another potential confounding factor in the original study is that RHD patients have long been known to have difficulty in reading emotions. Such a deficit could account for many of their social impairments (Heilman, Blonder, Bowers, & Crucian, 2000; Hailman, Bowers, Speedie, & Coslett, 1984). While Happé et al. (1999) balanced emotional expression and facial information across the ToM and non-ToM materials, it may be that reading emotions was less critical for understanding the control items, the humour of which was based on physical cause and effect. To examine this possibility, cartoons in the current study were classified in terms of degree of emotion (high versus low) experienced by a character in each cartoon.

Finally, in addition to a group analysis, we carried out an individual analysis paying special attention to lesion locus. Historically, imaging studies have had the benefit of lesion studies to point to regions of particular interest in a given domain. Yet the ascendance of neuroimaging now affords lesion studies the opportunity to follow their lead in determining regions of possible functional significance in variousdomains.

METHOD

Participants

A total of 11 right-hemisphere damaged (RHD) participants (6 male, 5 female) between the ages of 44 and 76 (mean = 61) took part in the study. Education levels ranged from 12 to 19 years (mean = 14). The RHD group was recruited through Healthsouth Braintree Rehabilitation Network, Braintree, MA. All of the patients were right-handed and all with the exception of LL were native English speakers. Each participant suffered unilateral right-hemisphere damage due to cerebro-vascular accident (CVA) as confirmed by either MRI or CT scan (see Figure 2). See Table 1A and 1B for patient profiles. A total of 20 non-brain-damaged age- and education-matched control (NC) participants (13 women and 7 men) took part in the study. Their ages ranged from 49 to 79 (mean = 66). Education levels ranged from 12 to 20 years (mean = 15). All were free of past or present diagnosis of psychiatric disorders, history of drug or alcohol abuse, and developmental or learning disorders.

Materials and procedure

A total of 30 single-frame cartoons were taken from popular magazines, newspapers, and books (e.g., *New Yorker*, *Far Side*). There were three types

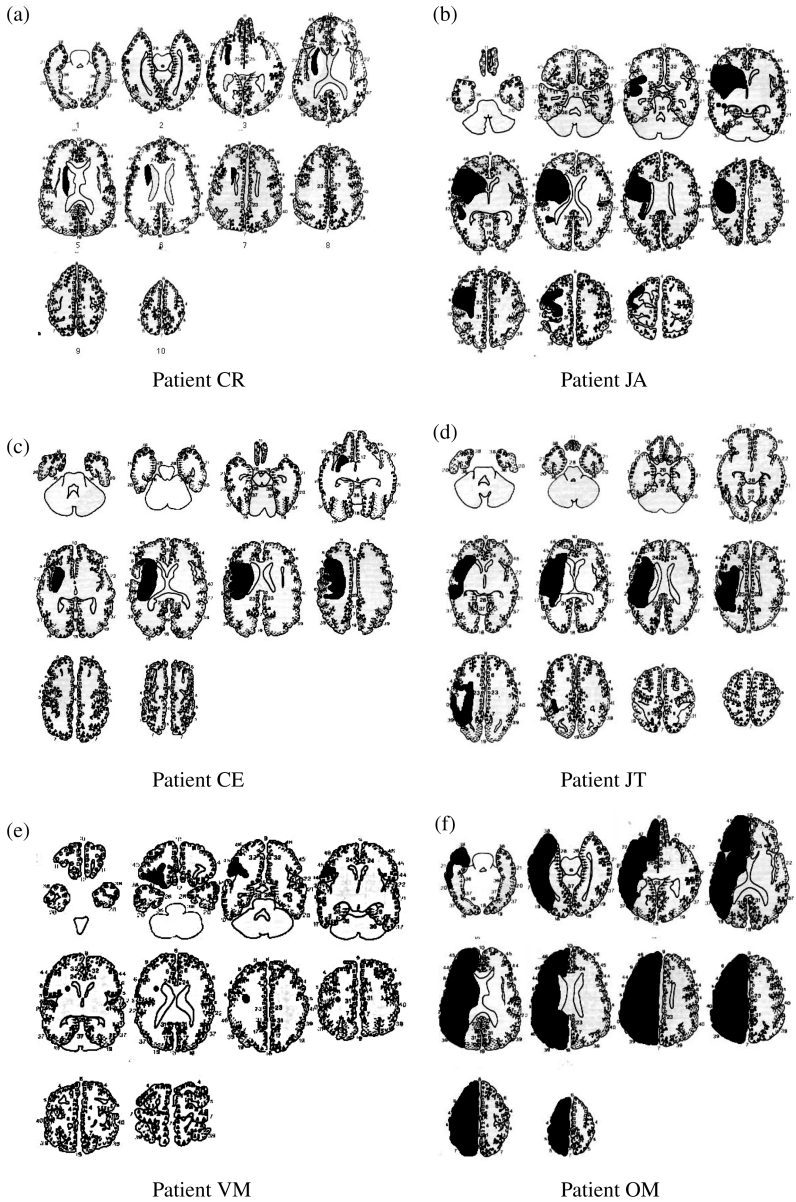

Figure 2. Patient scans. The right hemisphere is depicted on the left side.

TABLE 1A
Patient characteristics

Patients	*Age at onset*	*Age at testing*	*Lesion location*	*Brodmann areas*
MM	62	68	F P C W	1, 2, 3, 12, 44, BG, INS
CR			F T P W	BG, INS
CE	54	58	F W	1, 2, 3, 4, 6, 22, 40, 44, INS
JA	63	66	F T P C W	1, 2, 3, 4, 5, 6, 9, 22, 44, BG
EB	54	57	UK	
BF	43	44	T C W	21, 22, 37,38
OM	52	54	Pr F T P C W	Entire hemisphere except occipital corex; including BG, Thal, INS, Amyg
LL	63	64	UK	
JT	63	64	T P C W	1, 2, 3, 6, 21, 22, 39, 40, 44, INS
CF	75	76	T W	Thal (?)
VM	68	69	Pr F C W	45, 47

Abbreviations: Pr = Prefrontal, F = Frontal, P = Parietal, T = Temporal, C = Cortex W = White matter, BG = Basal ganglia, INS = Insula, Thal = Thalamus, Amyg = Amygdala.

of cartoons, with 10 instances of each type, that were distinguished by the requirements for appreciating the humour. Each item contained at least two characters and contained a similar number of visual elements. *Physical* cartoons required an inference about prior physical events but not about mental states. *First-order theory of mind* cartoons required assessment of the ignorance or false belief of at least one of the characters. *Second-order theory of mind* cartoons required awareness of either deception or fooling in which one character was aware of what the other character knew or did not know, and was taking advantage of this knowledge.

The 10 instances of each cartoon type were further divided into high- and low-emotion categories. Emotionality was defined on the basis of ratings obtained from a set of 20 pilot participants (undergraduate students) who evaluated a larger set of 40 cartoons. The pilot participants rated a specified

TABLE 1B
Lesion characteristics

Patient	*Lesion description*
CR	Small basal ganglia and central deep white matter lesion
CE	Large posterior deep white matter lesion
JA	Large dorsolateral frontal and basal ganglia lesion
BF	Moderate lateral middle temporal lesion
OM	Very large fronto temporo parietal lesion
JT	Large fronto temporal lesion
CF	Small white matter lesion in temporal isthmus
VM	Small lateral convexity lesion

character's emotion on a scale of 1 to 4 in which 1 indicated not at all emotional, 2 slightly emotional, 3 emotional, and 4 very emotional. Cartoons classified as high emotion had a mean rating of at least 3.2 (range 3.2–4); those classified as low emotion had a mean rating of 2.7 or below (range 1–2.6). The cartoons rated between 2.6 and 3.2 during piloting were eliminated from the final set, in order to maintain a clear distinction between high- and low-emotion items.

For all but one task, the original versions of the 30 cartoons were used. In the humour-rating task, stimuli consisted of each original cartoon paired with a non-humorous foil, created by altering or deleting one element. (See Figure 3 for an example.)

Participants were tested individually. Testing required either one or two sessions lasting from 45 to 90 minutes each, depending on the preference of

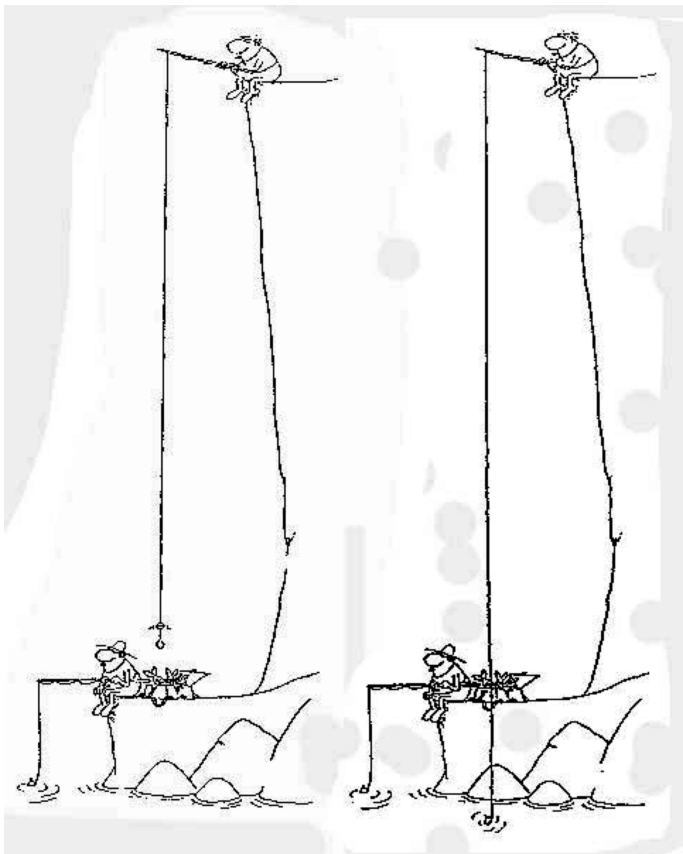

Figure 3. Real (left) and foil second-order theory of mind items.

the participant. Tasks were administered to all participants in the same order, as described below.

Humour-rating task. The members of each cartoon pair, the original and the altered foil, were presented side by side, with left–right order of correct/foil counterbalanced as far as possible across items within each condition. The experimenter pointed out each important element in each cartoon, as well as the element that was different or missing across the two cartoons in the pair. Participants were then instructed to concentrate on the cartoon on their left (the other was removed from view) and to rate it for funniness on a 4-point scale with verbal labels: 1 = not at all funny, 2 = slightly funny, 3 = funny, 4 = very funny. After the participant rated the first cartoon, the procedure was repeated for the second cartoon in the pair. Ratings were recorded by the experimenter. If the participant gave the same rating to both the original and the foil, he or she was asked which cartoon they found funnier. If the participant did not think either was funnier, he or she was asked to indicate the cartoon that "most people would find funnier". At the end of the task, the foils were removed from view.

Emotion-rating task. The participant was shown each of the original cartoons in isolation. For each cartoon, the experimenter pointed to a particular character and asked the participant to rate that character's emotion on a 4-point scale: 1 = not at all emotional, 2 = slightly emotional, 3 = emotional, 4 = very emotional.

ToM and non-ToM inference task. For each cartoon, participants were asked two different types of ToM question and two types of non-mentalistic control questions. The questions were asked in the same order for all items in order to make the task less confusing for patients. Participants were told at the outset that some of the questions would not be relevant to some cartoons. Participants were told directly, "Don't be afraid to say 'no' to some of these questions because often the best answer is 'no'." Answers to the inference questions were tape-recorded at the time of testing and later coded.

ToM questions. These questions, together, tapped participants' facility with mentalistic analysis regardless of the type of cartoon. *Knowledge* questions were designed to tap the participant's ability to attribute first-order knowledge. The question "Is there anything that X doesn't know, or is unaware of, that is *important* to the humour of this cartoon?" was posed for each of two characters in each cartoon. If the participant answered "yes", he or she was then asked "What is it that X doesn't know or is unaware of?" Participants were told directly at the outset that they should provide "yes"

answers if and only if a cartoon character's lack of awareness was important to the humour of the cartoon. *Deception questions* were designed to assess participants' second-order intentional attribution ability. The questions "Is X trying to trick or fool Y?" and "Is Y trying to trick or fool X?" were posed about the two characters in each cartoon.

Non-ToM questions. Two non-mentalistic questions that require interpretation and inference by participants provided an approximate control for the inferential difficulty and abstractness of the ToM questions. *Danger* questions were posed about each of the two characters in each item: "Is X in danger?" and "Is Y in danger?" "Yes" answers were followed by "What is X (or Y) in danger of?" A single *Impossibility* question was posed for each item: "Is there anything impossible going on in this cartoon?" Participants were told at the outset that "unlikely/improbable" and "impossible" were not the same for the purposes of the study, and that they should reserve the term "impossible" for events that were "absolutely impossible" and not simply unlikely or improbable. If the participants answered "yes", they were then asked to specify what was "impossible".

Trails A & B: Executive function (set-shifting). Trails A & B assesses set-shifting abilities associated with executive function and is sensitive to general cognitive decline (Spreen & Strauss, 1998). In Trails A, a respondent must use a pencil to trace in sequence letters from A to Z that are arrayed in a random configuration on a single sheet of paper. In Trails B, the respondent must trace a sequence that alternates between letters and numbers, i.e., A-1-B-2-C-3 etc. Respondents were instructed at the beginning of the test to correct any errors they noticed while working. Times to the nearest second were measured with a stopwatch and recorded for both Trails A and B.

Homographs task: Central coherence. A homograph-reading task was used to assess a participant's sensitivity to verbal context. The test consisted of 40 sentences, each containing a homograph with both a rare and frequent interpretation. The homograph in a sentence could either be preceded or followed by disambiguating context. The total set of 40 included equal numbers (10) of the following four types: low-frequency pronunciation with target word placed *before* disambiguating context; low-frequency pronunciation with target word placed *after* disambiguating context; high-frequency pronunciation with target word placed *before* disambiguating context; high-frequency pronunciation with target word placed *after* disambiguating sentence context. The sentences were presented in a booklet, with one sentence per page. The sentences were shuffled and presented in a pseudo-random order. Half of the participants were presented with the initial shuffle while the other half were presented with the reverse order of the initial shuffle.

We followed the procedure of Frith and Snowling (1983), Happé (1997), and Joliffe and Baron-Cohen (1999), and did not forewarn the participants that they would be reading homographs; no practice was given, and sufficient reading level and vocabulary proficiency were assumed. Nothing occurred during the procedure to place doubt on these assumptions except for one RHD patient, whose data were excluded. Participants were instructed to read sentences aloud at a normal pace. They were told that if they made a mistake, they should correct the mistaken word or phrase. After each sentence was read, either the experimenter or the participant turned the page for the next sentence, depending on the participant's preference. Correct readings, corrections, and mistakes were recorded. The dependent measure was the number of times a participant used the context-inappropriate pronunciation. If a participant used the inappropriate pronunciation initially but then corrected it, he or she was given credit for a correct response. Examples of homographs can be found in the Appendix.

RESULTS

Emotion perception task

Participants rated cartoon characters' emotion on a scale from 1 (not at all emotional) to 4 (very emotional). The primary purpose of the ratings was to assess whether both the stroke patients and the controls were comparably sensitive to the distinction between high and low levels of emotion conveyed in cartoon humour, and the secondary purpose was to confirm that the stroke patients were capable of using a 4-point rating scale. For analysis, each participant's ratings were averaged across the five trials within each cell of the design. Thus, each participant contributed a total of six data points representing 2 levels of Emotionality (high, low) $\times 3$ levels of Cartoon Type (physical, first-order, second-order). Statistical analysis was based on mixed design analysis of variance (ANOVA) with group (RHD, NC) as the sole between-subjects factor.

Both patients and control participants made a highly reliable distinction between the high- and low-emotion items, thereby indicating sensitivity to the manipulation, $F(1, 29) = 446.776$, $p < .001$, $MSE = .206$. Overall, patients rated the stimuli as less emotional, $F(1, 29) = 4.439$, $p = .0439$, $MSE = .388$. However, most important is that the data contained no indication of a deficit in perceiving the difference in emotion. The patients' emotionality effect (high emotionality mean = 3.70, low emotionality mean = 2.12) was as strong or stronger than the effect found for the controls (high emotionality mean = 3.79, low emotionality mean = 2.43). The interaction of Group $\times$ Emotionality was not reliable, $F(1, 29) = 2.540$, $p = .1218$, $MSE = .206$. In summary, both controls and patients

were sensitive to the different levels of emotionality, and as a group the patients were able to use a 4-point rating procedure reliably. The data contained no indication of a deficit bearing on perception of emotional intensity in these cartoons.

Humour-rating task

Participants also rated each original cartoon and the corresponding altered item on a scale from 1 (not at all funny) to 4 (very funny). Each participant's ratings were again averaged across the five trials within each cell of the design. Thus, each participant contributed a total of 12 data points representing 2 levels of Cartoon Version (original, altered foil) $\times 2$ levels of Emotionality (high, low) $\times 3$ levels of Cartoon Type (physical, first-order, second-order). Statistical analysis was based on a mixed design analysis of variance (ANOVA) with Group (RHD, NC) as the between-subjects factor.

As expected, participants rated the original cartoons as funnier, $F(1, 29) = 221.187$, $p < .0001$, $MSE = .535$, but the controls made a more pronounced distinction (original mean = 2.95, altered mean = 1.46) than did the patients (original mean = 2.30, altered mean = 1.44), $F(1, 29) = 15.788$, $p = .0004$, $MSE = .535$. The two groups treated the altered foils similarly, but the patients did not increase their ratings as much for the original versions. Thus the RHD patients' ratings suggest a reduced appreciation of humour. Consistent with the pattern of means within the interaction, the data also contained a less relevant, reliable difference between the two groups' overall funniness ratings. Furthermore, the patients' decreased appreciation of humour was apparent in each cartoon type (physical, first-order, second-order). When the data were analysed separately for each type, the Group $\times$ Version interaction was uniformly reliable. Thus the RHD group's diminished appreciation of humour is not related to a ToM impairment in any obvious way.

The emotionality of items also had an impact, though the impact appeared to be analogous for controls and patients. High emotionality items (mean = 2.10) were generally rated as funnier than low emotionality items (mean = 1.97), $F(1,29) = 23.081$, $p < .0001$, $MSE = .068$. When the data for the three cartoon types were analysed separately, the emotionality effect was reliable for the physical and first-order items and represented a statistical trend for the second-order items. None of these analyses revealed a reliable Group $\times$ Emotionality interaction, $F < 1.0$ for all three cartoon types. As we discussed earlier, the study was in part designed to examine the emotionality factor. It is noteworthy that a decreased sensitivity to emotion does not offer a compelling explanation for either the theory of mind or humour deficits observed.

Theory of mind and non-ToM inference questions

Answers to the inference questions were coded as representing the following response types: Hits, which were correct and relevant responses; Hit-Minus, which were plausible but not entirely relevant responses; False Alarms, which were implausible and irrelevant responses; Misses which indicate that a relevant answer was missed; and Correct Rejections, which were appropriate negative ("no") responses. Explanations were not required following responses to the "Deception" questions, and hence no Hit-Minuses were scored for these. All responses were coded independently by two researchers. Good inter-rater reliability (98%) was achieved. Disagreements were resolved through discussion. Sample responses can be found in the Appendix.

Primary analysis: Combined ToM and non-ToM questions. To obtain the most stable measures of overall performance, composite scores were calculated for each question: Knowledge, Deception, Danger, and Impossible. Hits and Correct Rejections were scored as 1, Hit-Minuses as 0, and Misses and False alarms were scored as −1. An individual composite score for a particular stimulus item was the sum of these values from the questions asked about the two characters in a cartoon. Averages for the stimuli in each cell of the design were used as the dependent measure for analysis.

The basic finding is that the RHD patient group showed a selective deficit in ToM. To characterise the overall findings in summary form, we first averaged together the scores for the two kinds of ToM questions (knowledge and deception) and those for the two kinds of non-ToM questions (danger and impossibility), and then carried out a mixed-design ANOVA that included Group (RHD, NC), Emotionality (high, low), and Question Type (ToM, non-ToM) as factors. The critical result is an interaction involving Group × Question Type, $F(1, 29) = 6.846$, $MSE = .019$, $p = .0140$. The performance difference between patients and controls was much more pronounced for the ToM questions (NC mean = .61, RHD mean = .43) than for the nonToM questions (NC mean = .31, RHD mean = .27).

Other reliable effects included the following. The controls (mean = .46) performed marginally better overall than the RHD patients (mean = .35), $F(1, 29) = 4.104$, $MSE = .084$, $p = .0521$. Both groups performed better on the ToM than the non-ToM questions, $F(1, 29) = 83.852$, $MSE = .019$, $p < .0001$. The overall ease of the ToM questions strongly suggests that the selective RHD deficit was *not* simply a scaling artifact (that an impaired group will show greater deficits on harder tasks): an artifact of nonspecific complexity or difficulty is unlikely to account for the RHD patients' ToM deficit.

Finally, the data contained a highly reliable main effect of Emotionality and a highly reliable interaction between Emotionality and Question Type: high-emotionality items were responded to more correctly than low-emotionality items, and this difference was more pronounced for ToM items (high mean = .61, low mean = .44) than for non-ToM items (high mean = .33, low mean = .24), $F(1, 29) = 5.743$, $MSE = .008$, $p = .0232$. The controls and patients, however, treated the emotionality factor analogously, $F < 1.0$ for all interactions involving group and emotion.

Subsidiary analyses of ToM and non-ToM performances. In light of the significant effects observed in the composite measures, we next performed subsidiary ANOVAs for each of the four Question Type (knowledge, deception, impossibility, danger) and Cartoon Type (physical, first-order, second-order) combination to explore the source of the reliable effects observed above. These secondary analyses each included Group and Emotionality as factors. The results of these nine ANOVAs are summarised in Table 2. It can be seen that the second-order items, which relied most

TABLE 2
Individual question ANOVA results

		Cartoon type		
Question		*Physical*	*First-order*	*Second-order*
Knowledge	RHD mean	.305	.432	.009
	NC mean	.507	.528	.323
	$F(1, 29) =$	2.718	1.367	5.598
	MSE	.215	.095	.249
	p value	.110	.252	.025*
Deception	RHD mean	.645	.667	.545
	NC mean	.825	.785	.700
	$F(1, 29) =$	6.005	2.719	4.880
	MSE	.076	.073	.069
	p value	.021*	.110	.035*
Danger	RHD mean	.004	.195	.043
	NC mean	.023	.238	.089
	$F(1, 29) =$	0.862	2.060	2.128
	MSE	.012	.012	.014
	p value	.361	.162	.155
Impossibility	RHD mean	.455	.482	.423
	NC mean	.480	.515	.504
	$F(1, 29) =$	0.035	.064	.674
	MSE	.265	.243	.138
	p value	.853	.802	.418

* indicates significant result

extensively on a sophisticated ToM capacity, produced significant differences between the control and RHD groups for the knowledge and deception questions, and no differences between groups for the danger and impossibility questions. In addition, both ToM questions produced reliable group differences. Comparable analyses for the first-order and physical items showed that the danger and impossibility questions never yielded reliable group differences. The only other effect of note comes from the deception question on the physical items. This was due to several False Alarms from the RHD group, where they claimed to detect deceit when there was none.

The emotion factor figured in only one reliable interaction for the knowledge question data for the physical items: The high-emotionality items were harder for the patients (high mean = .25, low mean = .36) but easier for the controls (high mean = .60, low mean = .42), $F(1, 29) = 5.194$, $MSE = .063$, $p = .030$. Comparable effects did not appear elsewhere in the data, and we have no ready interpretation for this exceptional effect tied to emotionality.

In sum, these analyses suggest a tentative restriction of the scope of impairment: RHD patients' deficit appeared to be carried by their difficulties when a second-order inference is required, either by virtue of the cartoon type or by virtue of the question asked about a cartoon. Furthermore, RHD patients' difficulty was not due to an inability to articulate the relevant features of the stimulus materials, since patients performed significantly worse than normal controls on forced-choice (deception) questions (which do not require verbal explanation) as well as on open-ended (knowledge) ToM questions about the second-order items.

Trails A & B

Data from two RHD patients were not available for Trails due to reading difficulties associated with neglect or visual field defects. While it was possible to effectively direct these patients' attention to the relevant features in the cartoon and homographs tasks, directing a patient's attention in Trails would amount to doing the task for them. Trails times were entered into a mixed design ANOVA. Overall, participants took much longer to complete Trails B than Trails A, $F(1, 27) = 37.872$, $MSE = 2136.580$, $p < .0001$, and the RHD patients were slower overall (mean = 130.3 s) than the controls (mean = 72.0 s), $F(1, 27) = 10.064$, $MSE = 4185.614$, $p = .0037$. However, the test did not distinguish between the groups in the anticipated fashion. The RHD patients showed more of an increase from Trails A (mean = 77.7 s) to Trails B (mean = 182.9 s) than the controls (Trails A mean = 43.9 s, Trails B mean = 100.2 s); but the critical interaction of Group × Condition was only marginally reliable,

$F(1, 27) = 3.484$, $MSE = 2136.580$, $p = .0720$. Weak as it was, this effect was driven by one RHD patient, BF, whose score was more than two standard deviations above the RHD mean. With BF excluded from the analysis, the interaction of Group × Condition was lost, $F(1, 26) = 1.176$, $p = .2880$.

In what follows, we use the difference between patients' Trails B and A as a measure of general impairment, which includes components of executive function, to examine the nature of the differences between the control and RHD patient groups. We first carried out a multiple regression analysis with the average *ToM* question scores as a dependent measure and with Group (as a dummy coded variable) and the Trails B–A index as the two independent variables. The overall proportion of variance accounted for ($R^2 = .229$) was significant, $F(2, 26) = 3.854$, $p = .0342$. Group membership was a reliable predictor of ToM performance, $t = 2.131$, $p = .0427$, but Trails B–A was not, $t = -.954$, $p = .3490$. Thus, the overall difference in ToM performance between the RHD and control group could not be accounted for in terms of the Trails measure of executive flexibility and/or general impairment.

A second regression included average *non-ToM* question performance as the dependent measure and the same two independent variables, Trails B–A and Group. The overall proportion of variance accounted for ($R^2 = .222$) was again reliable, $F(2, 26) = 3.702$, $p = .0385$. However, in this analysis, the group variable was not reliable, $t = .499$, $p = .6219$, while the Trails B–A measure was significant, $t = -2.349$, $p = .0267$. In contrast to the ToM regression analysis, the non-ToM composite scores were clearly related to Trails performance, and group membership did not seem to have any independent role in accounting for performance. It seems that the difference between our participant groups is more appropriately viewed in terms of mentalising ability. The group distinction on ToM can be separated from group differences in general levels of test performance or other pre-existing differences, while the group distinction on non-ToM inferences questions cannot be separated from overall declines in test performance or pre-existing differences.

Homographs

The homographs test proved quite easy for both the RHD and control populations, resulting in a ceiling effect. The median and modal response in the (easier) after-context condition was 0.0 for both the RHD and control groups. In the more difficult before-context condition, the controls' median and modal score was still 0.00, and the RHD group was not reliably different from the controls. The good performance of the RHD patients suggests that

TABLE 3
Individual RHD patient summaries

Patient initials	*Trails*	*ToM cartoons*	*Non-ToM cartoons*
OM	1.9607	+0.6460	−0.9860
EB	−0.7501	+0.3446	+0.4354
MM	+0.8432	−0.3299	−1.1076
JT	–	−0.6262	+0.9057
CF	−0.4190	−0.8152	+0.0180
VM	−0.8122	−0.9481	+0.7049
CE	–	−1.3824	+0.5305
BF	+5.3129	−2.2305	−1.7417
LL	+2.1883	−1.7962	−1.2133
JA	+0.8225	−2.3583	+0.2135
CR	−0.0259	−2.4656	−0.7271

RHD patients *z* scores based on the NC mean and *SD*.

Higher scores on cartoons reflect better performance; lower scores on Trails reflect better performance (quicker shifting of set). A value of 0.00 indicates the mean performance for the control group.

their central coherence abilities are not as impaired as those with high-functioning autism or Asperger's syndrome, and places doubt as to any relationship between weak central coherence and ToM, at least for the group analysis of the RHD patients tested here. The ceiling effect excluded the possibility of any direct comparisons to our other measures.

RHD lesion location

We identify patients who showed the greatest impairment of ToM ability as a basis for speculating on functional–anatomical relations. We computed a "z score" for each patient based on the mean and standard deviation from the 20 non-brain-damaged control participants. Thus, a score of 0.00 indicates that a patient scored at the mean for the control group. For each patient we calculated one z score for ToM performance and another z score for non-ToM performance. The z scores are displayed in Table 3.

Selective ToM deficit. As a basis for exploring lesion effects, we examined the patients who showed the worst relative performance on ToM, with "worst" defined as showing the greatest difference between ToMz score and non-ToM z score in Table 2. These were patients JA, JT, VM, CE, and CR. The most extreme dissociation was found with patient JA. These patients all had (premotor) frontal lesions and all had white matter lesions; only VM's damage included prefrontal cortex. Four of the five had damage

to the inferior frontal cortex (BA 44/45), and three of five had damage to BA 6, an area of interest from the functional imaging work. It should be noted, however, that these three patients (CE, JA, JT) had very large lesions. Also of note, three of five had insula, basal ganglia, and somatosensory (BA 1, 2, 3) damage. Two additional patients, BF and LL, did very poorly on both ToM and non-ToM tasks. Lesion information was not available for LL. BF's lesion included the superior temporal sulcus (a region of interest) as well as inferior portions of the temporal pole, which has extensive connections to the amygdala and related limbic regions. See Table 1 for more detailed neurological data and Figure 2 for scans of notable patients.

Despite these intriguing commonalities, however, there is always at least one patient who serves as an exception to the rule. Patient MM performed slightly better at the ToM than the non-TOM items, and OM, who had damage to all of these areas, showed the opposite profile: a selective sparing of ToM abilities. Patient OM will be discussed in more detail below.

Examination of individual patients and their associated lesions is consistent with a couple of tentative conclusions. First, the neuroimaging literature on ToM does not as yet predict areas that, when damaged, will selectively affect ToM performance. Though group-level analyses show a selective ToM deficit (particularly on second-order ToM and the ability to detect deception), individual analyses reveal several profiles within the group: such as first-order ToM difficulties, deficits on all measures, sparing on all measures, and, in the case of OM, selective sparing of ToM. The RH areas that best predicted ToM impairment in our sample were inferior frontal cortex (BA 44, 45) perhaps in conjunction with insular/somatosensory cortex. The functional significance of the regions will be discussed below.

DISCUSSION

We presented 11 RHD patients and 20 non-brain-damaged controls with tasks measuring non-ToM and graded ToM inferential abilities. The RHD patients differed from the non-brain-damaged controls on a composite measure of mentalistic attribution ability. The patients differed most clearly from controls in the ability to attribute second-order intentional states; that is, in attributing knowledge about knowledge and the ability to detect deception. They did not differ reliably from the controls in their attributions of first-order intentional states or on the non-ToM inferential measures. The data contained several other effects that reinforce the dissociation of mentalistic from non-mentalistic ability. First, there were no clear-cut differences between the groups on the Trails measure that taps overall level of function and set-shifting capabilities associated with executive function. The strong correlation between performance on the control questions and

the Trails measure, and the lack of a strong correlation with the ToM questions, suggests that at least certain kinds of non-ToM reasoning do indeed rely on general cognitive processes, and that ToM reasoning ability, on the other hand, can remain independent. It is noteworthy, however, that the two patients who scored lowest on the ToM measures also performed very poorly on Trails, suggesting that these aspects of EF must be relatively intact for successful ToM reasoning. Correlations between mentalistic and non-mentalistic reasoning and central coherence abilities were not performed due to a ceiling effect on the homographs task. Yet the fact that RHD patients did so well on this task suggests that a significant deficit in central coherence is unlikely to account for RHD patients' difficulties with ToM.

The RHD patients' ratings of the emotions of the cartoon characters were indistinguishable from those of the control group, effectively ruling out an emotion-perception deficit as a compelling source for the observed ToM impairment. This study also replicates findings that RHD patients have difficulty appreciating humour (e.g., Bihrle, Brownell, Powelson, & Gardner, 1986; Shammi & Stuss, 1999). We note, however, that this difficulty was not restricted to the ToM items but extended to the physical items as well, and that the RHD patients' good performance on the first-order items suggests that the humour in the items did not prevent the patients from understanding them. This diminished appreciation of humour extended to all the cartoon types and hence is not apparently due to a ToM deficit. In addition, the humour appreciation effect was a subtle one. The RHD patients were slightly less likely than controls to rate the real cartoons as funnier than their foil counterparts, often rating them as equally funny. However, when a patient was forced to choose between a foil and an original cartoon that he or she had given identical ratings, the patient was often able to choose the real over the foil cartoon.

Finally, the similarity of emotionality ratings between the RHD and NC groups suggests that deficits in the perception of emotion cannot account for diminished mentalising capacity in the RHD patient group. The link between RHD and emotion deficits is beyond debate. Our point in the present context is that an impairment in perceiving emotion is not a likely account for the ToM results observed in this study.

The claims concerning the most likely anatomical substrate of ToM must remain tentative due to our small sample size and the nature of our data. The ToM imaging studies, most notably those by Gallagher et al. (2000) and Brunet et al. (2000) who also used wordless cartoons, provide hypotheses as to which RH regions may be involved in our tasks of mental state attribution. These areas include: the medial frontal cortex/paracingulate cortex, the right middle frontal gyrus, superior temporal sulcus, temporo-parietal junction, and the precuneus. Our patients had damage to all the

commonly described areas in various combinations, although lesion locus did not predict performance on the ToM items. It is therefore possible, and quite probable, that higher-order cognitive functions such as ToM are embedded in a distributed neural network, rather than dedicated and localisable cell assemblies.

All of the patients showing the strongest dissociation between ToM and non-ToM had premotor frontal lesions, and only VM had prefrontal damage. Moreover, all of them had damage to underlying white matter. The areas that best predicted a ToM deficit were inferior frontal (BA 44/45), right insula, and somatosensory cortices (BA 1, 2, 3). All of the patients showing selective deficits had damage to these areas in some combination. Four out of the five patients of note had damage to BA 44/45, three out of five had damage to BA 6, three out of five had insula damage, and three out of five had somatosensory damage. Patient CR was the only patient in this profile who did not have damage to BA 44/45, although her small lesion (basal ganglia, insula) is closely connected to these regions. Three out of the five also had basal ganglia damage. The only other patient who had damage to all of these areas was OM. Patient OM is an exceptional case and will be discussed in more detail below.

The one area of overlap with the nonverbal ToM imaging studies was BA 6 in the middle/medial frontal gyrus. Of the four patients with damage to this area, three of them had selective difficulty with the ToM items. This area has also been activated in an imaging study of metaphor processing (Bottini et al. 1994), and possibly reflects its role in enabling alternative interpretations, either through the inclusion of contextual factors or weaker semantic associations (Beeman, 1998).

The imaging studies of ToM have not singled out BA 44/45 as particularly important in this domain. BA 44/45 is the right hemisphere homologue to Broca's area, and has been implicated in ToM by Rizzolatti, Gallese, and colleagues, as part of a "mirror' system of mindreading (Rizzolatti, Fadiga, Gallese & Fogassi, 1996; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). This system is claimed to contain "mirror" neurons, which fire not only when an individual executes a particular action, but also when an individual observes that action carried out by another. The implications of such a system for learning, imitation, and even empathy (via strong limbic connections) are far reaching, and fit closely with a mechanistic account of ToM called "simulationism" (Gallese & Goldman, 1998). It is noteworthy that the Broca's aphasia patients tested by Happé et al. (1999), some of whom had damage to the same area in the left hemisphere, did not show a similar deficit.

The function of the insula, basal ganglia, and somatosensory cortices in these tasks is not yet clear, although Bunge, Dudukovic, Thoamson, Vaidya, and Gabriele (2002) found right insular (and inferior frontal cortex)

activation in a task requiring subjects to disregard distractors while finding a target, and Wyland, Carrie, Kelley, Macrae, Gordon, and Heatherton (2003) noted significant insula activation during tasks requiring thought suppression. The insula has also been implicated in the suppression of overt motor behaviour and task switching (Dove, Pollmann, Schubert, Wiggins, & von Cramon, 2000; Garavan, Ross, & Stein, 1999; Rubia, Taylor, Smith, Oksannen, Overmeyer, & Newman, 2001). In addition to its role in inhibitory processes, the insula has been implicated in somatosensory integration (Augustine, 1996), and may play a special role in the integration of limbic and cortical contributions to decision making, particularly in the social domain, i.e., "somatic marking" (Damasio, 1994; Damasio, Tranel, & Damasio, 1991). Indeed, in a recent study of emotional and social intelligence, Bar-On, Tranel, Denburg, and Bechara (2003) found that patients ($n = 3$) with right insular/somatosensory damage were impaired in social intelligence relative to executive function and measures of general intelligence, and relative to patients with damage to areas other than those involved with somatic marking (e.g., other than orbitofronal/ventromedial cortices, amygdala, insular/somatosensory cortices). Although their measures were not direct tests of ToM and relied heavily on self-reports, our findings lend support to their hypothesis of right insular/somatosensory involvement in social cognition, perhaps via its role in somatic marking.

The insula and basal ganglia have also been implicated in the recognition and expression of disgust, a social emotion thought to co-opt phylogenetically older substrates for food aversion (Calder, Keane, Manes, Antoun, & Young, 2000; Rozin & Fallon, 1987). These regions have increasingly been implicated in higher cognitive functions (Middleton & Strick, 2000) and have been suggested to play a causal role in autism spectrum disorders (Bishop, 1993; Damasio & Maurer, 1978; Maurer & Damasio, 1982).

We noted above that the patients all had white matter damage. This may not be surprising considering that white matter damage is common in stroke patients, particularly following middle cerebral artery infarction. Although all of our RHD participants had white matter damage in one form or another, the location and extent of the damage may be significant.

Proportionately, there is more white matter in the RH than the left. The heavily myelinated axons that comprise white matter are more efficient at relaying information over longer distances, whereas the densely packed cell bodies characteristic of the left hemisphere are more efficient at local, serial, processing (Galaburda, 1995; Gur et al.,1980). The RH has more high-level "associative" (integrative) cortex, and less cortex dedicated to specific motor or sensory functions, and hence requires a mechanism for more distant, rapid transmission of information. The effect of white matter damage on ToM computations may again be due to the ineffective integration of a multi-step, multi-regional computation, either within or between the hemi-

spheres. This "connectivity" argument has also been put forward regarding high-functioning autism and Asperger's syndrome (Minshew, Goldstein, & Siegel, 1997), with imaging studies showing deficits in functional connectivity compared to controls (Castelli, Frith, Happé & Frith, 2002; Just, Cherkassky, Keller, & Minshew, 2004).

This study also reveals the importance of analysing both group and individual performances. While our group results are fairly straightforward, there are several interesting individual profiles, some of which reflect the group profile, some of which do not. It is assumed that normal functioning is subserved by very similar neural substrates in different subjects, and while this may be true in adults for lower-level functions such as primary auditory and visual processing, it is not clear whether this holds for higher-order functions such as ToM. The same function might be mediated by different network configurations in different subjects, hence accounting for some of the variability in the functional imaging studies in normals (see Figure 1). While physiological processing efficiencies will likely place a limit on the number of possible configurations, the number is quite likely to be greater than one. In addition, regarding lesion studies, the neural instantiation of a (partly) recovered function may bear little resemblance to the original network that subserved that function. It is well known that motor recovery following stroke, for example, is accomplished not only by incorporating tissue adjacent to the damaged area but also often involving homologous areas in the contralesional hemisphere (Dijkhuizen et al., 2001; Gerloff, Altenmuller, & Dichgans, 1996; Nelles et al., 1999; Rijntjes & Weiller, 2002). It is probable that the kind, location, and extent of lesion will determine the available avenues for neuronal "re-assemblage" to re-establish the affected function. As a consequence of this, different lesions may not only affect the immediate dysfunction in the network, but also the evolution of network reorganisation in the course of recovery. Hence, a subject with a relatively small lesion may "recover" to an inefficient level of functioning that could involve bilateral representations, while a subject with a much larger lesion might reorganise in a much more efficient way by utilising completely contralesional representations. Variability in "wild-type" neuronal substrates and differences in reorganisation during recovery might both provide structural explanations of the peculiar lesion vs function distribution in our population, particularly OM, who had a massive right hemisphere lesion but normal ToM performance.

In sum, these results reflect those of Bird, Castelli, Malik, Frith, and Husain (2004), whose patient GT performed well on ToM measures despite having a large bilateral medial prefrontal lesion., an area consistently activated in ToM imaging studies (Frith & Frith, 1999; Gallagher & Frith, 2003), and suggest that neuroimaging findings alone are insufficient to understand the functional anatomy of mental state attribution. Our results

support and in many ways refine the original findings of Happé et al. (1999), and extend those of Bar-On et al. (2003). Future researchers may want to include a more sensitive measure of coherence, tests of additional EF components, as well as alternative ToM measures, in order to disambiguate this issue of task demands from domain-specific reasoning more broadly.

Manuscript received 4 May 2005
Revised manuscript received 10 October 2005
PrEview proof published online 10 April 2006

REFERENCES

Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews*, *22*, 229–244.

Bar-On, R., Tranel, D., Denburg, N. L., & Bechara, A. (2003). Exploring the neurological substrate of emotional and social intelligence. *Brain*, *126*, 1790–1800.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Baron-Cohen, S., Ring, H., Moriarty, J., Schmitz, B., Costa, D., & Ell, P. (1994). Recognition of mental state terms. Clinical findings in children with autism and a functional neuroimaging study of normal adults. *British Journal of Psychiatry*, *165*, 640–649.

Beeman, M. (1998). Coarse semantic coding and discourse comprehension. In M. Beeman & C. Chiarello (Eds.), *Right hemisphere language comprehension: Perspectives from cognitive neuroscience*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Benowitz, L. I., Moya, K. L., & Levine, D. N. (1990). Impaired verbal reasoning and constructional apraxia in subjects with right hemisphere damage. *Neuropsychologia*, *28*, 231–241.

Bihrle, A. M., Brownell, H. H., Powelson, J. A., & Gardner, H. (1986). Comprehension of humourous and nonhumourous materials by left and right brain-damaged patients. *Brain and Cognition*, *5*, 399–411.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain*, *127*, 914–928.

Bishop, D. V. (1993). Annotation: Autism, executive functions, and theory of mind. A neuropsychological perspective. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *34*, 279–293.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., et al. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain*, *117*, 1241–1253.

Brownell, H., Griffin, R., Winner, E., Friedman, O., & Happé, F. (2000). Cerebral lateralization and theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition* (pp. 306–333). Oxford, UK: Oxford University Press.

Brunet, E., Sarfati1, Y., Hardy-Baylé, M-C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage*, *11*, 157–166.

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabriele, J. D. E. (2002). Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron*, *33*, 301–311.

Calarge, C., Andreasen, N. C., & O'Leary, D. S. (2003). Visualizing how one brain understands another: A PET study of theory of mind. *American Journal of Psychiatry*, *160*(11), 1954–1964.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of digust following brain injury. *Nature Neuroscience*, *3*, 1077–1078.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*, 1839–1849.

Channon, S., & Crawford, S. (2000). The effects of anterior lesions on performance on a story comprehension test: Impairment on a theory of mind-type task. *Neuropsychologia*, *38*(7), 1006–1017.

Damasio, A., & Maurer, R. (1978). A neurological model for childhood autism. *Archives of Neurology*, *35*, 777–786.

Damasio, A., Tranel, D., & Damasio, H. (1991). Somatic markers and the guidance of behaviour: Theory and preliminary testing. In H. Levin, H. Eisenberg, & A. Benton (Eds.), *Frontal lobe and dysfunction* (pp. 217–229). New York: Oxford University Press.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam.

Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books, Inc.

Dijkhuizen, R. M., Ren, J., Mandeville, J. B., Wu, O., Ozdag, F. M., Moskowitz, M. A., et al. (2001). Functional magnetic resonance imaging of reorganisation in rat brain after stroke. *Proceedings of the National Academy of Science*, *98*(22), 12766–12771.

Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., & von Cramon, D. Y. (2000). Prefrontal cortex activation in task switching: An event-related fMRI study. *Cognitive Brain Research*, *9*(1), 103–109.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R.J., Frackowiak, R.S.J., et al. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*, 109–128.

Frith, C., & Frith, U. (1999). Interacting minds: A biological basis. *Science*, *286*, 1692–1695.

Frith, U. (1989). *Autism: Explaining the enigma*. Oxford, UK: Basil Blackwell.

Frith, U., & Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *British Journal of Developmental Psychology*, *1*, 329–342.

Galaburda, A. (1995). Anatomic basis of cerebral dominance. In R. J. Davidson & K. Hughdahl (Eds.), *Brain asymmetry* (pp. 51–73). Cambridge, MA: MIT Press.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind". *Trends in Cognitive Sciences*, *7*, 77–83.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of the mind' in verbal and nonverbal tasks. *Neuropsychologia*, *38*, 11–21.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.

Gallese, V., & Goldman, A. (1998). Mirro neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493–501.

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings for the National Academy of Sciences of the USA*, *96*(14), 8301–8306.

Gerloff, C., Altenmuller, E., & Dichgans, J. (1996). Disintegration and reorganisation of cortical motor processing in two patients with cerebellar stroke. *Electroencephalography and Clinical Neurophysiology*, *98*(1), 59–68.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport*, *6*, 1741–1746.

Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 83–116). New York: Cambridge University Press.

Gur, I. K., Packer, J. P., Hungerbuhler, J. P., Reivich, M., Orbist, W. D., Amarnek, W. S., et al. (1980). Differences in the distibution of gray and white matter in human cerebral hemispheres. *Science*, *207*, 1226–1228.

Happé, F. (1997). Central coherence and theory of mind in autism: Reading homographs in context. *British Journal of Developmental Psychology*, *15*, 1–12.

Happé, F. (2000). Parts and wholes, meaning and minds: Central coherence and its relation to theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism and developmental cognitive neuroscience*. Oxford, UK: Oxford University Press.

Happé, F., Brownell, H., & Winner, E. (1999). Acquired "theory of mind" impairments following stroke. *Cognition*, *70*, 211–240.

Happé, F., Ehlers, S., Fletcher, P., Frith, U., Johansson, M., Gillberg, C., et al. (1996). 'Theory of mind' in the brain. Evidence from a PET scan study of Asperger syndrome. *Neuroreport*, *8*, 197–201.

Heilman, K. M., Blonder, L. X., Bowers, D., & Crucian, G. P. (2000). Neurological disorders and emotional dysfunction. In J. Borod (Ed.), *The neuropsychology of emotion* (pp. 367–412). New York: Oxford University Press.

Heilman, K. M., Bowers, D., Speedie, L., & Coslett, H. B. (1984). Comprehension of affective and nonaffective prosody. *Neurology*, *34*, 917–921.

Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioural and Brain Sciences*, *21*, 101–148.

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in Cognitive Science*, *8*, 25–32.

Jarrold, C., Butler, D.W., Cottington, E.M., & Jimenez, F. (2000). Linking theory of mind and central coherence bias in autism and in the general population. *Developmental Psychology*, *36*, 126–138.

Jolliffe, T., & Baron-Cohen, S. (1997). Are people with autism or Asperger's Syndrome faster than normal on the Embedded Figures Task? *Journal of Child Psychology and Psychiatry*, *38*, 527–534.

Jolliffe, T-D., & Baron-Cohen, S. (1999). Linguistic processing in high-functioning adults with autism or Asperger syndrome: Is local coherence impaired? *Cognition*, *71*, 149–185.

Just, M. A., Cherkassky, V. L., Keller, T. A., & Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: Evidence of underconnectivity. *Brain*, *127*, 1811–1821.

Maurer, R. G., & Damasio, A. R. (1982). Childhood autism from the point of view of behavioural neurology. *Journal of Autism and Developmental Disorders*, *12*, 195–205.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia output and cognition: Evidence from anatomical, behavioural and clinical studies. *Brain and Cognition*, *42*, 183–200.

Minshew, N. J., Goldstein, G., & Siegel, D. (1997). Neuropsychologic functioning in autism: Profile of a complex information processing disorder. *Journal of the International Neuropsychological Society*, *3*, 303–316.

Nelles, G., Spiekermann, G., Jueptner, M., Leonhardt, G., Muller, S., Gerhard, H., et al. (1999). Reorganisation of sensory and motor systems in hemiplegic stroke patients. A positron emission tomography study. *Stroke*, *30*(8), 1510–1516.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive function deficits in high-functioning autistic individuals: relationship to theory of mind. *Journal of Child Psychology and Psychiatry*, *32*, 1081–1105.

Parkin, A. J., & Lawrence, A. (1994). A dissociation in the relation between memory tasks and frontal lobe tests in the normal elderly. *Neuropsychologia*, *32*(12), 1523–1532.

Pennington, B., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *37*, 51–87.

Pennington, B., Rogers, S., Bennetto, L., Griffith, E., Reed, D., & Shyu, V. (1997). Validity test of the executive dysfunction hypothesis of autism. In J. Russell (Ed.), *Executive functioning in autism*. Oxford, UK: Oxford University Press.

Povinelli, D. (2000). *Folk physics for apes*. Cambridge, MA: MIT Press.

Rijntjes, M., & Weiller, C. (2002). Recovery of motor and language abilities after stroke: The contribution of functional imaging. *Progress in Neurobiology*, *66*(2), 109–122.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

Rowe, A., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). 'Theory of mind' impairments and their relationship to executive function following frontal lobe excisions. *Brain*, *124*, 600–616.

Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological Review*, *94*, 23–41.

Rubia, K., Taylor, E., Smith, A. B., Oksannen, H., Overmeyer, S., & Newman, S. (2001). Neuropsychological analyses of impulsiveness in childhood hyperactivity. *British Journal of Psychiatry*, *179*, 138–143.

Russell, J. (1997). How executive disorders can bring about an inadequate theory of mind. In J. Russell (Ed.), *Autism as an executive disorder*. Oxford, UK: Oxford University Press.

Sabbagh, M. (1999). Communicative intentions and language: Evidence from right-hemisphere damage and autism. *Brain and Language*, *70*, 29–69.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*, 1835–1842.

Shah, A., & Frith, U. (1983). An islet of ability in autism: A research note. *Journal of Child Psychology and Psychiatry*, *24*, 613–620.

Shah, A., & Frith, U. (1993). Why do autistic individuals show superior performance on the block design task? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *34*, 1351–1364.

Shallice, T. (2001). 'Theory of mind' and the prefrontal cortex. *Brain*, *124*, 247–248.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterisation of empathy deficits following prefrontal brain damage: The role of the right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience*, *15*, 324–337.

Shammi, P., & Stuss, D. T. (1999). Humour appreciation: A role of the right frontal lobe. *Brain*, *122*, 657–666.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.

Springer, S., & Deutsch, G. (1998). *Left brain, right brain: Perspectives from cognitive neuroscience* (5th ed.). New York: W.H. Freeman & Company.

Stone, V., Baron-Cohen, S., & Knight, K. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, *10*, 640–656.

Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for 'theory of mind'. *Brain*, *124*, 279–286.

Tomasello, M. (1998). Uniquely primate, uniquely human. *Developmental Science*, *1*, 1–16.

Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, *10*, 723–726.

Wyland, Carrie L., Kelley, W. M., Macrae, C.N., Gordon, H. L., & Heatherton, T. F. (2003). Neural correlates of thought suppression. *Neuropsychologia*, *41*(14), 1863–1867.

Zaitchik, D., Koff, E., Brownell, H., Winner, E., & Albert, M. (2004). Inference of mental states in patients with Alzheimer's disease. *Cognitive Neuropsychiatry*, *9*, 301–313.

APPENDIX: EXAMPLES OF RESPONSE CODING

Sample responses for second-order theory of mind cartoon (Figure 3)

1. Knowledge question

Hit: "He [bottom man] doesn't know he's about to get his fish stolen."
Hit-Minus: "He [bottom man] doesn't know there's another fisherman."
Correct Rejection: "No" [top man]
Miss: "No" [bottom man]
False Alarm: "He [top man] doesn't know his line isn't in the water."

2. *Deception question*

Hit: "Yes" [top man]
Miss: "No" [top man]
Correct Rejection: "No" [bottom man]
False Alarm: "Yes" [bottom man]

3. *Danger and Impossibility questions*

It was determined that "Correct Rejections" should be given for "no" responses to the Danger and Impossibility questions, as nothing is physically impossible and it is quite reasonable to say that no one is in danger. However, several subjects answered that the top man was in danger of falling and that the bottom man was in danger of being hit by the hook, for instance. As these are reasonable answers, they were scored as Hit-Minuses.

Sample responses for first-order low-emotion cartoon

1. *Knowledge question*

Hit: "He doesn't know that a dog's not really in the hose."
Hit-Minus: "[the man doesn't know] where the dog is."
Miss: "No" (attributed to the man)
Correct Rejection: "No" (attributed to the dog)
False Alarm: "He [the dog] doesn't know where the mouse is."

2. *Deception question*

Correct rejection: "No."
False Alarm: "Yes"

3. *Danger & Impossibility questions*

There is no danger and nothing is impossible in this scenario. Any attributions of danger or impossibility were scored as False Alarms.

Sample responses for physical high-emotion cartoon

1. *Knowledge question*

Hits: [none available]
Hit-Minuses: "He unaware of how he's gonna get down."
Correct Rejection: "No"
Miss: [none available]
False Alarm: " He [man on bug] doesn't know that he dropped his net."

2. *Deception question*

As there is no deception in this item, all "No" responses were scored as Correct Rejections. Any attributions of deception ("Yes") were scored as False Alarms.

3. *Danger and Impossibility questions*

Attributions of danger to the man on the bug (e.g., of falling) were coded as Hits while attributions of danger to the other character were coded as False Alarms. Hits for the Impossibility question were given for reference to the size of the bug/butterfly, or for reference to the impossibility of catching such a large bug in a small net.

Homograph examples

1. The lead car finished in record time. (lower frequency/before context)
2. The lead in the box made it very heavy. (higher frequency/before context)
3. If the street maps are confusing, you can hire a guide to lead you around. (lower frequency/ after context)
4. Due to the weight of the lead pipes, the workers used a crane. (higher frequency/after context)

5. Wind and reset the watch before midnight. (lower frequency/before context)
6. Wind and sun weathered the farmer's skin. (higher frequency/before context)
7. The blind watchmaker refused to wind his most precious clock. (lower frequency/ after context)
8. The clouds moved in quickly as the wind picked up speed. (higher frequency/after context)