4CAPS: An Adaptive Architecture for Human Information Processing

Sashank Varma

Stanford Center For Innovations in Learning
450 Serra Mall, Building 160
Stanford, CA 94305-2055
sashank@stanford.edu

Marcel Just

Center for Cognitive Brain Imaging
Carnegie Mellon University
Pittsburgh, PA 15213
just+@cmu.edu

Abstract

4CAPS is a cognitive architecture of interest to both cognitive science and AI. 4CAPS is of interest to cognitive science because it supports models of neuroimaging data collected using fMRI and PET. 4CAPS should be of interest to AI because it organizes human information processing in an optimal and mathematically tractable form. This paper focuses on the *adaptivity* of 4CAPS models in the face of changing task demands and fluctuating resource availability. It illustrates this adaptivity in the domains of problem solving, spatial reasoning, and sentence comprehension. It also identifies new forms of adaptivity ripe for future research.

Introduction

Cognitive science and artificial intelligence have drifted apart over the past twenty years. The split was not deliberate, but rather the result of diverging interests. Cognitive science is currently in the midst of a neuroscience revolution. Neuroimaging studies of normal young adults and patients with focal lesions have yielded data that demand different computational explanations than those available in AI. Conversely, AI has been in the midst of its own revolution (or maturation) from free-form computational explorations to rigorous mathematical and statistical developments.

We contend that it is time for cognitive science and artificial intelligence to reconcile. This paper describes one form that this reconciliation can take: the 4CAPS cognitive neuroarchitecture. As a cognitive architecture, 4CAPS is a unified theory of high-level cognition cast not as an axiomatic theory, but rather as a computational formalism. But it is more than that. It is capable of accounting for neuroimaging data collected from both normal young adults and patients with focal lesions. As a computational formalism, 4CAPS should be of interest to AI researchers because it makes the interesting – and mathematically tractable – claim that human information processing is, at both the cognitive and cortical levels, *adaptive*. Specifically, it is an optimal solution to the problem of scheduling computation in a way that respects resource constraints and maximizes throughput.

After describing the operating principles embodied in 4CAPS, this paper examines two forms of adaptivity using as examples models of three domains of high-level cognition: problem solving, spatial reasoning, and sentence comprehension. The paper concludes with a consideration of other forms of adaptivity suggested by 4CAPS.

Operating Principles

4CAPS embodies a number of operating principles that govern cortical and cognitive information processing. (See Just and Varma (2005) for more details.)

Centers

A 4CAPS model consists of multiple *centers*, each corresponding to a distinct brain area (i.e., gyrus or sulcus). The number of centers is denoted M. Each center is a hybrid symbolic-connectionist computational system. From a symbolic perspective, it is a production system. Production rules encode procedural knowledge and declarative elements encode declarative knowledge. From a connectionist perspective, production rules are graded, declarative elements are annotated with activation levels, and the control structure is fully parallel. The thresholds on the condition elements of productions and the weights on their action elements cause them to function like the links of connectionist networks. The activations of declarative elements, which represent their relevance for current and future processing, function like the activation levels of the units of connectionist networks. A third connectionist property is the adoption of parallel processing: at each time step, all matching productions fire (i.e., there is no conflict resolution scheme), directing activation from declarative elements on their condition sides to declarative elements on their action sides modulo their weights. In this way, 4CAPS centers combine the computational power of variable binding over symbolic structures and the activation dynamics of connectionist networks.

Multiple Specialization

Each center can perform multiple *functions*. A function is a convenient abstraction for a set of productions and declarative elements that together implement a cognitively interesting operation, such as parsing or maintaining a goal stack. Each center can perform multiple functions, and conversely, each function can be performed by multiple centers. This proposal is intermediate between modularity, which maps one and only function to each center, and equipotentiality, which maps every function to every center. The *specialization* of a center *i* for a function *j* is denoted S_{ij}, where $S_{ij} \in [1, \infty)$. A value of 1.0 represents perfect specialization – performing one unit of function *j* requires 1.0 units of center *i*'s activation resources. Larger values represent lesser specializations. If there are *N* functions, then the total resource demand of center *i* at a particular point in time is:

$$\sum_{j=1}^{N} \left(A_{ij} \cdot S_{ij} \right)$$

where A_{ij} represents the units of function *j* performed by center *i*.

Resource Constraints

Each center possesses a finite supply of resources (i.e., activation) reflecting biological limitations on computation. Specifically, the resource capacity of center *i* is denoted C_i and the constraint:

$$\sum_{j=1}^{N} \left(A_{ij} \cdot S_{ij} \right) \le C_i \qquad (1)$$

is enforced at all times. Performing a sufficiently complex task can exhaust the resource supply of a center specialized for required functions. When this happens, processing spills over from well-specialized centers experiencing resource shortfalls to less-specialized centers with spare resources. The spillover of function is a natural consequence of the algorithms described below.

Capacity Utilization

The *capacity utilization* of a center *i* is the proportion of resources currently in use:

$$CU_i := \frac{\sum_{j=1}^{N} \left(A_{ij} \cdot S_{ij} \right)}{C_i}$$

CU is a measure of how hard a center is working. The fundamental claim of 4CAPS is that the CU of a center predicts neural activity in the corresponding brain area (i.e., activation as measured by fMRI or PET).

Collaborative Processing

A 4CAPS model consists of multiple centers that collaborate in a large-scale network to perform a task. The membership and topology of the network change over time. Two factors determine the degree to which a center participates in the network at a particular point in time. The first is whether it is specialized for any of the functions awaiting execution. The second is whether the center possesses spare resources to fuel their execution. The *allocation problem* is to assign functions to centers in a way that maximizes cognitive throughput while minimizing resource consumption.

More precisely, at each point in time, a subset of the *N* functions awaits execution. The activation requested by function *j* is denoted R_j. Recall that A_{ij} denote the amount of function *j* to be performed by center *i*; the goal is to determine their values. The A_{ij} are constrained in two ways. First, *M* constraints, one for each center *i*, stipulate that no center can perform more work than its resources allow; these are the (1) above. Second, *N* constraints, one for each function *j*, stipulate that as much of each function as possible (but no more than the requested amount) is performed:

$$\sum_{i=1}^{M} A_{ij} \le R_j \qquad (2)$$

Many different allocations satisfy the constraints (1) and (2), such as $A_{ij}=0$ for all *i* and *j*. We therefore require a measure of the goodness of an allocation. This measure can be expressed as a linear combination of the A_{ij} to be maximized:

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \left(W_{ij} \cdot A_{ij} \right) \qquad (3)$$

Defining the weights as $W_{ij} := 1/S_{ij}$ ensures that, resource considerations aside, function *j* will be assigned to the center *i* most specialized for it (i.e., whose S_{ij} is minimal).

(1), (2), and (3) constitute a linear programming (LP) problem that can be solved using the simplex algorithm (Dantzig & Thapa, 1997). The result is an allocation of functions to centers that maximizes cognitive throughput while minimizing resource consumption.

Adaptivity

A task can be decomposed into a partial order of functions to be performed. At any point in time, then, there exists a mixture of pending functions. Centers assemble themselves into a large-scale network capable of performing these pending functions. The membership and topology of the network – its nodal centers and the edges along which they exchange partial products – are adaptive in two critical ways. First, they adapt to changes over time in the mixture of functions to be performed. Second, they

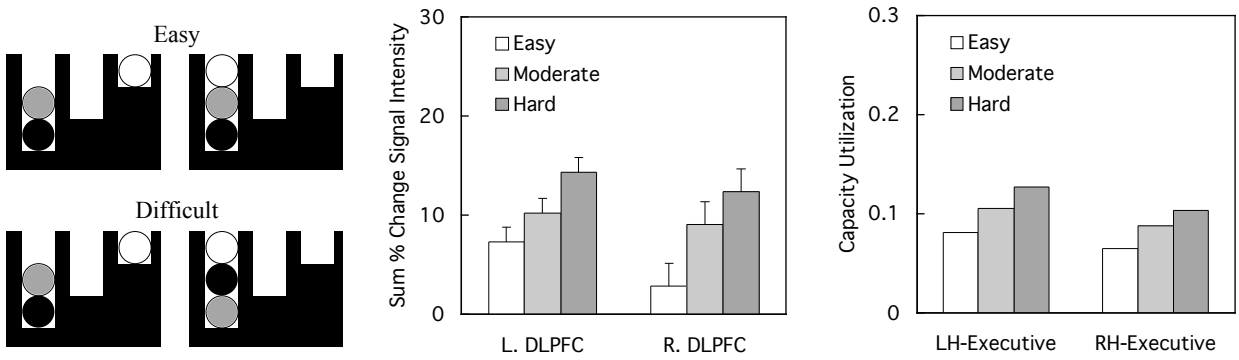

Figure 1. (a) Sample TOL problems. (b) fMRI data: frontal activations (Newman et al., 2003). (c) Model: Executive CUs.

adapt to changes in resource availability, both over the relatively short timescale of task performance and over the longer timescale of lifespan development. This section illustrates both kinds of adaptivity with 4CAPS models of various tasks (Just & Varma, 2005).

Adaptivity to Changing Task Demands

During the course of task performance, the mixture of pending functions to be performed can change. This can impose increasing resource demands on centers that already belong to the large-scale network.

For example, consider a 4CAPS model of Tower of London (TOL) problem solving. The TOL task is a variant of the Tower of Hanoi task that is easier to administer to neuropsychological patients. Example TOL problems are shown in Figure 1a. The first is rather simple; it can be solved by perceptual (i.e., similarity-based) considerations alone. The second is more difficult. It requires the establishment of subgoals to clear blocking balls before deeply buried balls can be moved. There is evidence that the frontal lobe, specifically right and left dorsolateral prefrontal cortex (DLPFC), perform goal operations: to the former area is attributed planning through the articulation of goal-subgoal hierarchies, to the latter selection between goal-based moves proposed by right DLPFC and perceptually-based moves proposed by posterior brain areas. For example, patients with frontal lesions can solve simple TOL problems that require only perceptually-based moves, but have difficulty with problems that require the formulation of a plan and the execution of goal-based moves (Shallice, 1982). Newman et al. (2003) collected more direct evidence for this claim. They had participants solve blocks of problems that varied in difficulty (i.e., the amount of goal-based processing required). Using fMRI, they measured the activations in left and right DLPFC; these are shown in Figure 1b. A 4CAPS model of TOL problem solving was constructed that includes centers corresponding to these two brain areas. The RH-Executive center is specialized for formulating plans, i.e., for generating goals when problem solving fails. The LH-Executive center is specialized for selecting between the goal-based moves proposed by RH-Executive and the perceptually-based moves proposed by centers corresponding to the parietal areas thought to perform visuospatial processing. The CUs of the Executive centers are shown in Figure 1c. They are an increasing function of problem difficulty, mirroring the neuroimaging data ($r=0.96$, $p<0.01$).

Another example of adaptation to changing task demands is found in a 4CAPS model of spatial reasoning. In the classic mental rotation task shown in Figure 2a, participants must decide whether the figures are congruent to or mirror images of one another. The classic behavioral result is that reaction time is a linear function of the angular disparity between the two figures (Shepard & Metzler, 1971). This is typically interpreted to mean that participants rotate one figure until it is aligned with the other; the greater the required rotation, the longer the response time. Carpenter et al. (1999) found an analogous result using fMRI. They had participants perform mental rotation in the scanner. Figure 2b plots activations observed in left and right parietal cortex as a function of angular disparity. The former area is thought to subserve the generation and maintenance of visuospatial representations, the latter the direction of visuospatial attention. The longer the required rotation, the greater the resource demands on these areas to perform the visuospatial representation and attention functions for which they are specialized, and thus the greater the observed activations. A 4CAPS model of mental rotation was constructed that includes centers corresponding to left and right parietal cortex. LH-Spatial generates intermediate visuospatial representations at each step along the rotation path; the greater the angular disparity, the longer the rotation path, the more steps it contains, and therefore the more intermediate representations that must be generated. RH-Spatial proposes possible rotations along

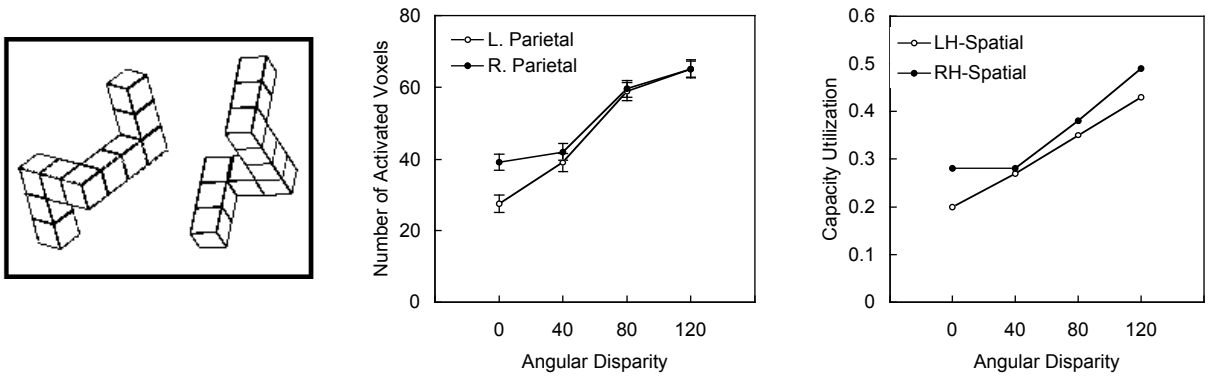

Figure 2. (a) Sample mental rotation problem. (b) fMRI data: parietal activations (Carpenter et al., 1999). (c) Model: Spatial CUs.

the *x*, *y*, and *z* axes at each step along the rotation path; the greater the angular disparity, the longer the rotation path, the more steps it contains, and therefore the more possible rotations that must be proposed. The CUs of the Spatial centers are shown in Figure 2c. They are an increasing function of angular disparity, mirroring the neuroimaging data ($r=0.87$, $0.10<p<0.15$).

Adaptivity to Changing Resource Demands and Availability

A second form of adaptivity is how 4CAPS handles changing resource demands. If the resource demands on a center well-specialized for functions to be performed are greater than the available resource supply, then processing will spill over to a center less-specialized for the functions, but possessing spare resources. We consider two variants of this form of adaptivity exhibited by a model of sentence comprehension.

The first form occurs when normal readers comprehend complex sentences. Consider the three sentences listed in Figure 3a. They are complex in that each consists of two clauses. The first simply conjoins the two clauses. Processing the two clauses is likely to tax the language network. The second and third sentences embed one clause (the 'relative clause') in the center of the other (the 'main clause'). Processing center-embedded sentences is especially resource-demanding because the partial products associated with the beginning of the main clause must be buffered while the interrupting relative clause is processed. The second sentence is a 'subject-relative' because the first noun phrase of the main clause is the subject of the relative clause; the third sentence is called an 'object-relative' because the first noun phrase of the main clause is the object of the relative clause. Object-relative are more resource demanding than subject-relatives because the first noun phrase of the main clause must be buffered longer before its position in the relative clause is reached (at which point it can be integrated into the emerging representation). Many behavioral studies find that object-relative sentences take longer to comprehend than subject-relative sentences, with the difference localized to the relative clause. Just et al. (1996) found an analogous result using fMRI: the more complex a sentence, the more activation observed in left inferior frontal gyrus (left IFG; Broca's area) and left

Conjoined Actives:	The senator attacked the reporter and admitted the error.
Subject Relative:	The senator that attacked the reporter admitted the error.
Object Relative:	The senator that the reporter attacked admitted the error.

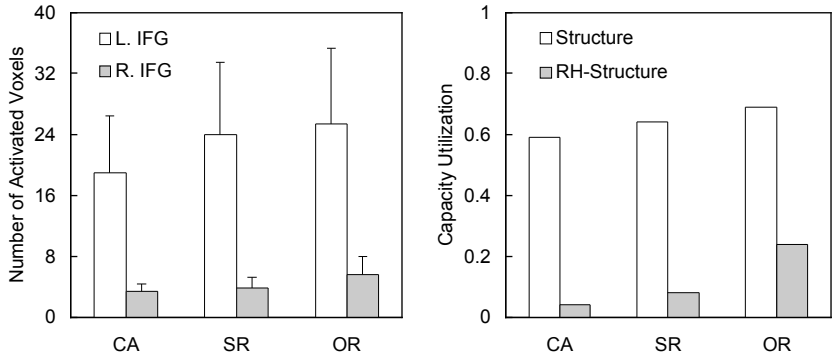

Figure 3. (a) Sentence types. (b) fMRI data: IFG activations (Just et al., 1996). (c) Model: Structure CUs.

superior/middle temporal gyrus (left S/MTG; Wernicke's area), the core areas of the language network. Critically, they also found that the more complex a sentence, the more likely it was to exhaust the resources available in the core areas and for processing to spill over to their right-hemisphere homologs, which are thought to be specialized for the same language functions, albeit at lower levels of efficiency. This is shown in Figure 3b for left and right IFG. A 4CAPS model of sentence comprehension was constructed with centers corresponding to left and right IFG and S/MTG. The Construct center corresponds to left IFG. It is specialized for the manufacturing of new structured representations (e.g., phrases) and for the maintenance of existing structured representations (e.g., verbal working memory). RH-Construct is specialized for the same functions, but to a lesser degree. As the complexity of a sentence increases, so does its buffering requirements, and therefore the demand on Construct's resources. As this demand exceeds the available supply at the most taxing portion of a sentence, processing spills over to RH-Construct. This adaptive recruitment of RH-Construct into the language network is shown in Figure 3c. The correlation between human and model performance is 0.98 (p<0.001).

The second form of adaptive recruitment occurs over longer timescales. Consider damage to a particular brain area, such as following stroke. The damage can be viewed as a drastic reduction in the available resource supply. Without adequate resources, the area will not be able to perform the functions for which it is specialized, and other areas specialized for the same functions will be recruited into the large-scale network on a more-or-less permanent basis. For example, Thulborn et al. (1999) imaged a patient who had suffered a lesion to left IFG but had recovered much of his language ability six months later. They had him read simple active sentences (e.g., *The senator attacked the reporter.*) and measured the activation in left and right IFG and S/MTG; these are shown in Figure 4a. The striking result is the spillover of processing from left to right IFG, especially in contrast to the left-lateralized activation in S/MTG. The 4CAPS sentence comprehension model was used to simulate these results. The Structure center was stripped of its resources to simulate the effect of the left IFG lesion. The CUs of the lesioned model's centers are shown in Figure 4b. The model correctly accounts for the shift in processing from the damaged center to its less-specialized though intact right-hemisphere homolog (r=0.99, p<0.01).

Conclusion

4CAPS bears directly on the fourth question of this symposium: *Is adaptiveness an essential component of intelligence?* 4CAPS places adaptivity at the heart of human information processing, in two ways. First, as the functional demands of task performance change, centers specialized for these functions are recruited into the large-scale network. Second, as the resource demands of task performance increase (or when the resource supply is drastically reduced following damage), processing spills over to other centers with similar functional specializations, recruiting them into the large-scale network.

4CAPS bears on two other questions of this symposium. The third question is: *Are multi-level heterogeneous approaches beneficial?* 4CAPS is multi-level in simultaneously addressing both the cognitive and cortical levels of human information processing. In fact, it makes the stronger claim that neither level can be completely understood without the constraint of the other. 4CAPS is heterogeneous in that its models are not monolithic systems, but rather mosaics of centers with interlocking functional specializations. The sixth question is: *Are specialized, modular components a reasonable approach to general intelligence?* 4CAPS claims that brain areas are not modules: they are not specialized for single functions and they do not interact in rigid ways. Modularity errs because it focuses on the nodes of large-scale networks

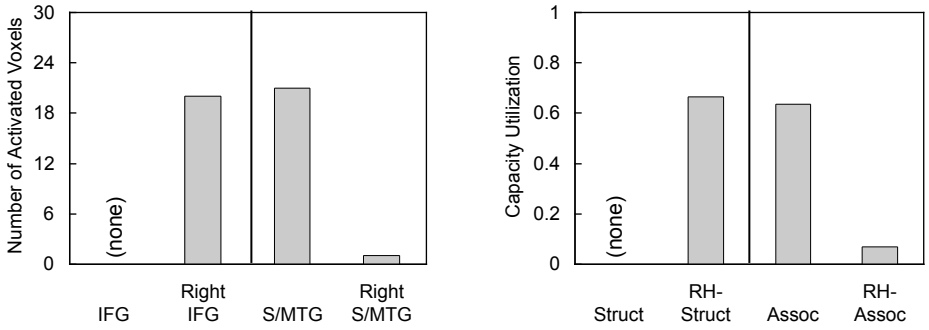

Figure 4. (a) fMRI data: normal and lesioned activations (Thulborn et al., 1999). (b) Model: normal and lesioned CUs.

and ignores the topologies that connect them: their initial assembly and dynamic reconfiguration with changing functional and resource demands.

We conclude by considering two implications of the adaptivity of 4CAPS for cognitive science and AI.

Fluid Intelligence and Induction

It is common to localize the flexibility of human cognition to particular brain areas. For example, Duncan et al. (2000) attribute *fluid intelligence* – the ability to solve abstract, culture-free problems – to prefrontal areas. 4CAPS suggests that the flexibility of human thought is not the product of particular brain areas, but rather the fluency with which they assemble themselves into large-scale networks appropriate for performing novel tasks. This suggests a novel view of *induction*. The problem of inducing new strategies has existed since at least Plato's *Meno*. The most common solution is to propose a learning mechanism. 4CAPS suggests a different solution: New strategies result in part from new organizations of existing brain areas into non-canonical large-scale networks. This is potentially very productive. For example, if there are M brain areas, then for a given non-empty subset containing m areas, there are $2^{m(m-1)/2}$ ways to connect them into a large-scale network. This large space of possible networks contains the canonical network for performing the task at hand as well as a number of alternate networks. 4CAPS provides a framework in which to think about how these alternatives are discovered. There are several circumstances under which a model might assemble its centers in a non-canonical way. One is a change in resource availability, whether transient (e.g., because of the removal of a typical load or the imposition of a new one) or permanent (e.g., through cognitive development, cognitive aging, following damage). When this happens, a non-canonical network might assemble, one that configures a novel subset of centers using a novel topology. The result of this change at the cortical level will be a new strategy at the cognitive level.

Linear Programming

A novel contribution of 4CAPS is the formalization of the adaptivity of human cognition using the LP formalism. The potential of this tool for modeling cortical and cognitive information processing is largely unexplored, and calls out for a joint effort between cognitive science and AI. For example, 4CAPS currently solves the allocation problem in a centralized manner: at each point in time, the resource demands of the functions to be performed and the constraints on resource availability are cast as an LP problem and simplex applied to determine an allocation of functions to centers. That this is currently done using centralized data structures and algorithms is an implementation detail. One area for future research is the development of methods for solving the allocation problem that are consistent with the parallel and distributed nature of cortical computation. Under such methods, each center will make allocation decisions, such as when to shift and when to take on excess processing, based primarily on locally-available information (i.e., from neighboring centers in the large-scale network). It is likely that these local allocation decisions will be suboptimal at the global level. A first step in this direction has already been taken by researchers interested in the solution of large LPs through parallel processing (Lustig & Rothberg, 1996; Maros & Mitra, 2000). We believe this to be fertile ground for future research at the intersection between cognitive science and AI (Boddy & Dean, 1994; Korf, 1990).

References

Boddy, M., and Dean, T. 1994. Decision-theoretic deliberation scheduling for time-constrained environments. *Artificial Intelligence* 67: 245-286.

Dantzig, G. B., and Thapa, M. N. 1997. *Linear programming 1: Introduction.* New York: Springer-Verlag.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F. N., and Emslie, H. 2000. A neural basis for general intelligence. *Science* 289: 457-460.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. 1996. Brain activation modulated by sentence comprehension. *Science* 274: 114-116.

Just, M. A., and Varma, S. 2005. *The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition.* Manuscript under review.

Korf, R. 1990. Real-time heuristic search. *Artificial Intelligence* 42: 189-211.

Lustig, I. J., and Rothberg, E. 1996. Gigaflops in linear programming. *Operations Research Letters* 18:157-165.

Maros, I., and Mitra, G. 2000. Investigating the sparse simplex algorithm on a distributed memory multiprocessor. *Parallel Computing* 26: 151-170.

Newman, S. D., Carpenter, P. A., Varma, S., and Just, M. A. 2003. Frontal and parietal participation in problem solving in the Tower of London: fMRI and computational modeling of planning and high-level perception. *Neuropsychologia* 41: 1668-1682.

Shallice, T. 1982. Specific impairments in planning. *Philosophical Transactions of the Royal Society London* B298: 199-209.

Shepard, R., and Metzler, J. 1971. Mental rotation of three-dimensional objects. *Science* 171: 701-703.

Thulborn, K. R., Carpenter, P. A., and Just, M. A. 1999. Plasticity of language-related brain function during recovery from stroke. *Stroke* 30: 749-754.