

Modelling the Supervisory System and Frontal Dysfunction: An Architecturally Grounded Model of the Wisconsin Card Sorting Task

Mariam R. Sood

(msood01@mail.bbk.ac.uk)

Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London WC1E 7HX, UK

Richard P. Cooper

(R.Cooper@bbk.ac.uk)

Abstract

We present a model of the Wisconsin Card Sorting Test, a classical neuropsychological test frequently used to assess deficits in executive functioning. The model is grounded in a cognitive architecture based on the Supervisory System theory of Norman and Shallice (1986) and evaluated against data from control subjects and several groups of neurological patients as reported by Stuss et al. (2000). The model is able to account for control performance across a range of dependent measures. When damaged in theoretically motivated ways it is also able to capture the behaviour of the different patient groups. Specifically, the model supports the association by Shallice et al. (2008) of the function of task-setting to left lateral prefrontal cortex, of the function of attentiveness to inferior medial prefrontal cortex, and of the function of monitoring to right lateral prefrontal cortex. The implication of these results for the supervisory system architecture and the localisation of function within prefrontal cortex are discussed.

Keywords: Cognitive architecture; Supervisory system; Wisconsin card sorting task; Frontal dysfunction.

Introduction

Several theories of the organisation of cognitive processes have been proposed over the last 25 years. These cognitive architectures generally comprise complex production systems, and normally have their roots in behaviours in specific cognitive domains (e.g., problem solving, as in, Soar: Newell, 1990; associative memory, as in ACT-R: Anderson, 2007; or immediate response tasks as in EPIC: Meyer & Kieras, 1997). While such architectures have been highly successful at accounting for a range of behavioural effects, they are not well suited to modelling the behaviour of neurological patients with focal brain damage. This is largely because it is unclear how the functional components of such architectures might be impaired without causing complete breakdown of the system. The cognitive architecture sketched by Norman and Shallice (1986) and elaborated by Shallice et al. (2008), in contrast, provides a modular view of cognition in which functional components may operate more or less efficiently, and hence neurological deficits might be more directly accounted for.

The Norman/Shallice theory draws a primary distinction between routine behaviour, which is generated by a lower level scheduling system – Contention Scheduling – and non-routine behaviour, which is effected by a higher level system – Supervisory System. This higher level system operates indirectly on behaviour by modulating the functioning of Contention Scheduling. When initially

described (Norman & Shallice, 1986), the situations requiring Supervisory System input were clearly enumerated but the subsystem's functioning was specified only in abstract terms. Those functions include what have since come to be known as executive functions such as task-setting, monitoring and working memory maintenance.

In a somewhat separate line of work, Shallice, Stuss and colleagues (e.g. Stuss et al., 2000; Shallice et al., 2008) have attempted to account for the deficits of several groups of patients with focal frontal lobe lesions in terms of deficits affecting specific executive functions which, they argue, are effected by different regions of the prefrontal cortex. Thus, the deficits of patients with left lateral prefrontal lesions across a range of tasks are interpreted as reflecting impaired task-setting, while the deficits of right lateral prefrontal patients are interpreted as reflecting impaired monitoring. Similarly, the deficits of patients with focal lesions affecting inferior medial prefrontal regions are interpreted as reflecting an impaired ability to sustain attention to a task, while the deficits of patients with focal lesions affecting superior medial prefrontal cortex are interpreted as reflecting an impairment in “energisation”, i.e., mobilisation of cognitive resources, corresponding phenomenologically to cognitive effort.

Shallice et al. (2008) relate the four executive functions discussed in the previous paragraph to the Supervisory System, with a specific focus to how the two accounts relate within a simple task-switching study. However these authors provide only an informal characterisation of the functions. They do not provide a precise computational instantiation of the ideas. The goal of this paper is to provide and evaluate such an instantiation. More specifically we present a computational account of the heterarchical organisation of the Supervisory System. The account is grounded in a model of a specific task – the Wisconsin Card Sorting Test (WCST). This widely used test of executive function provides multiple dependent measures that are sensitive to frontal lobe damage (Milner, 1963). We report simulations of the behaviour of control subjects and of four patient groups, comparing our results with those of Stuss et al. (2000), who tested patients and controls on the task.

The following sections briefly discuss the cognitive architecture in which the model is framed, the Wisconsin card sorting test and the neuropsychological group study that provides the target data. Following this, we present the model itself, the methodology for modelling control and

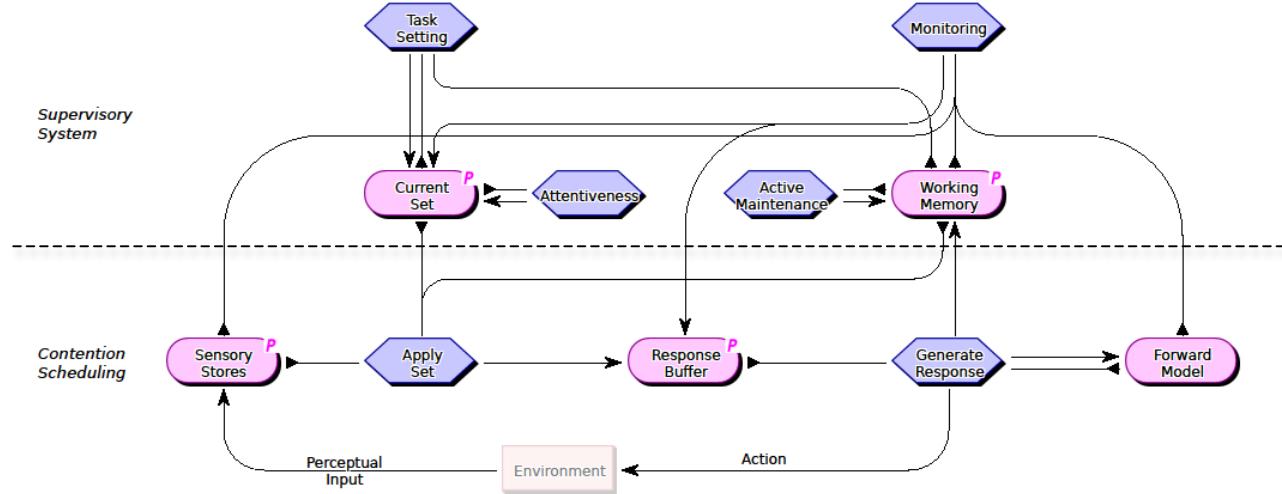


Figure 1: The proposed functional organisation of the Supervisory System architecture. Hexagonal boxes represent processes while rounded rectangles represent buffers or storage systems. Arrows show hypothesised connectivity between components.

patient performance, and the respective simulation results. We conclude by considering the implications of this work for the computational specification of the Supervisory System and more generally for the functional organisation of higher cognition.

The Supervisory System Architecture

The Supervisory System proposed by Shallice, Stuss and colleagues (e.g., Shallice et al., 2008) is a heterarchical system comprising, amongst other things, four core subprocesses: *task-setting*, *active monitoring*, *energisation* and *attentiveness*. The evidence for this organisation is drawn from neuropsychological case studies where the nature of deficits exhibited by frontal patients show subtle differences based on the lesion location. For example, the impairment exhibited by left prefrontal patients may be understood as resulting from inefficient task strategy formation while right prefrontal patients make errors that suggest poor ability to monitor internal and external events. The deficits of inferior medial prefrontal patients may stem from a characteristic lack of attention while superior medial prefrontal patients exhibit a longer (30%) start up delay in task execution compared to other groups (for a review, see Shallice & Cooper, 2011).

The cognitive architecture of the model described in this paper is derived from the Contention Scheduling / Supervisory System theory and is depicted in figure 1. Processing within the Contention Scheduling components of architecture is as follows: perceptual input enters *Sensory Stores*. Potential responses are generated from this by *Apply Set* subject to application of the current stimulus-response mapping set. These responses are passed to a *Response Buffer* before being generated as actions. The *Generate Response* process also maintains *Forward Model*, which represents the anticipated sensory feedback of the system's actions. The Supervisory System modulates the behaviour

of Contention Scheduling by two key processes: a) *Monitoring*, which compares sensory feedback with anticipated sensory feedback and rejects the current stimulus-response mapping if there is a mismatch (i.e., an unanticipated sensory input) by clearing *Current Set*, and b) *Task Setting*, which sets a stimulus-response mapping when *Current Set* is empty. Two other supervisory processes, *Attentiveness* and *Active Maintenance*, work to counteract decay which is assumed to operate on elements within *Current Set* and *Working Memory*. With the exception of *Energisation*, the model adequately represents all other subprocesses of the Supervisory System theory.

The Wisconsin Card Sorting Test

The Task

In order to evaluate the Supervisory System architecture we consider its application to a specific task: the Wisconsin Card Sorting Test (WCST). The WCST exists in various forms. The version simulated here is the 64A version used by Stuss et al. (2000). In this version of the task, subjects are required to sort a deck of cards, 64 in total presented one at a time, into four groups. Each card has a picture of a specific shape in variable numbers and colours (e.g., one red triangle or four blue squares; see figure 2). Four “target” cards, differing with respect to the number, colour and shape of items they depict, are provided and subjects are required to place each successive card from the main deck under one of the four target cards. In the 64A version, subjects are informed of the three possible sorting criteria – sort by colour, sort by number or sort by shape – prior to the test. After each card is sorted, the subject is given feedback. Based on the feedback, the subject should attempt to infer the correct sorting rule and use it for subsequent sorts. Once the subject correctly sorts 10 cards consecutively, the experimenter changes the rule without warning. The ideal

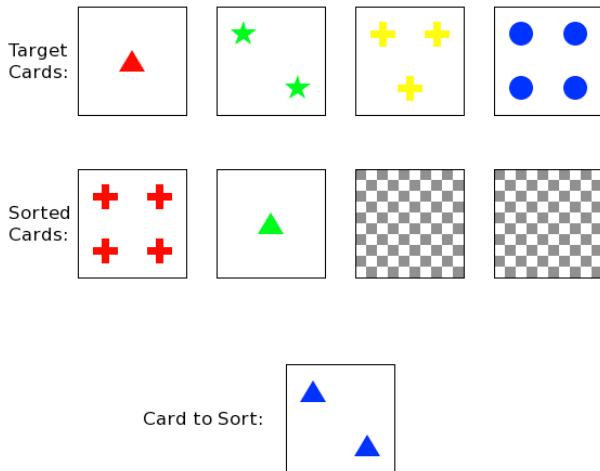


Figure 2: The Wisconsin Card Sorting Test, after two cards have been sorted according to the colour of their symbols and as a third card (two blue triangles) is presented for sorting.

subject will detect this and select a new rule, based on the feedback after each sorting attempt.

Neurologically healthy subjects have little difficulty in this task. However patients with frontal lesions are prone to perform poorly, frequently showing incapacity to change the rule when the feedback is negative, i.e. they tend to 'persevere', but also showing 'set loss' errors, where they appear to correctly infer a rule, but fail to follow that rule for ten consecutive sorting trials, even with positive feedback.

Neuropsychological Evidence

The motivation behind choosing the WCST for evaluation of the supervisory system architecture over other executive tasks is the availability of detailed empirical data published by Stuss et al. (2000) on patients categorised with focal lesions on the four brain regions of theoretical interest. The empirical study carried out by Stuss et al. (2000) tested seven groups of subjects. Four groups had focal frontal lesions on left/right dorsolateral prefrontal cortex (LDL/RDL), superior medial (SM) and inferior medial (IM) prefrontal regions. The fifth and the sixth patient groups had lesions affecting left/right non-frontal brain regions and the seventh group comprised neurologically healthy subjects. The subjects were tested on three versions of WCST: 128, 64A and 64B. In the 128 version, subjects were not provided any instructions on how to perform the test. In 64A version, subjects were informed of the three possible sorting criteria beforehand, while in 64B version subjects were also alerted when the rule was about to change. In each case, the errors made by subjects were classified into four categories: perseveration of preceding category (PPC: an incorrect response that matches the preceding sorting criterion), perseveration of preceding response (PPR: an incorrect response that matches exactly the features on the preceding trial), set-loss (an incorrect response following

three or more consecutively correct responses) and other errors. Patients with frontal lesions, compared to those with non-frontal lesions and controls, exhibited more PPC, PPR and set-loss errors. The error patterns exhibited by the different frontal groups revealed subtle differences. For example, in the 64A condition, all frontal groups except IM showed significantly more PPC and PPR errors than controls. In contrast, the IM group made significantly more set-loss errors than patients from other frontal groups.

Modelling the WCST

Model Assumptions and Description

The model discussed here is an elaboration of the heterarchical Supervisory System theory (figure 1), with its components configured for the WCST 64A condition of the empirical study by Stuss et al. (2000). Consider first the three buffers and two processes that make up Contention Scheduling. When a card is to be sorted, a propositional representation of the card appears in *Sensory Stores*. *Apply Set* then consults *Current Set* for a representation of the current sorting rule (e.g., sort by colour) and uses this in conjunction with the representation in *Sensory Stores* (e.g., two blue triangles) to generate a putative response (e.g., place the card on the right-most pile) which is stored in *Response Buffer*. *Generate Response* then produces the actual response (storing a copy in *Working Memory*), together with a representation of the anticipated consequences of the response – the *Forward Model*. (In the current implementation *Forward Model* is ignored, since the anticipated consequence of any action is positive feedback.)

Processing within the Contention Scheduling components is modulated by the Supervisory System components. First, *Monitoring* may detect negative feedback in the sensory store (or more generally, a mismatch between the contents of *Forward Model* and *Sensory Store*). In such situations, *Monitoring* will clear *Current Set* (on the assumption that the current sorting rule is inappropriate). Second, *Task Setting* may generate a putative sorting rule and place a representation of that rule in *Current Set*. This occurs when *Current Set* is empty (e.g., because the representation of the previous sorting rule in *Current Set* has either decayed or been explicitly deleted by *Monitoring*). Generation of a putative sorting rule depends on the contents of *Sensory Buffer* and recent responses stored in *Working Memory*.

Elements in the two supervisory buffers (*Current Set* and *Working Memory*) have activation values that decay over time. If the activation values fall below a threshold, the buffer contents cannot be accessed. The supervisory processes of *Attentiveness* and *Active Maintenance* work in direct opposition to decay, exciting buffer elements so as to prevent their loss.

The model's behaviour may be summarised as follows: At the beginning of the task, the first sorting schema is generated at random from among the three possible schemas: sort-by-colour, sort-by-number and sort-by-form. When a card is presented, the *Contention Scheduler* sorts

Table 1: Parameters of the model.

Parameter	Range	Description
Monitoring _{exogenous}	0 – 1	If impaired, feedback is not acted upon
Monitoring _{endogenous}	0 – 1	If impaired, drop in attention is not monitored
Taskset _{none}	0 – 1	If impaired, unable to switch strategy
Taskset _{random}	0 – 1	If impaired, unable to produce efficient strategy, a random strategy is chosen
Attentiveness _{persistence}	0 – 1	The persistence (decay) rate associated with current-set activation $activation_{new} = Attentiveness_{persistence} \times activation_{old} \pm noise$
Attentiveness _{boost}	1 – 10	Boost rate associated with current-set activation, set at 1.25 $activation_{new} = Attentiveness_{boost} \times activation_{old} \pm noise$
Attentiveness _{threshold}	0 – 1	Current-set activation threshold, set at 0.5
Memory _{persistence}	0 – 1	The persistence (decay) rate associated with working memory activation $activation_{new} = Memory_{persistence} \times activation_{old} \pm noise$
Memory _{boost}	1 – 10	Boost rate associated with working memory activation, set at 1.25 $activation_{new} = Memory_{boost} \times activation_{old} \pm noise$
Memory _{threshold}	0 – 1	Working memory activation threshold, set at 0.5

the card according to the sorting criterion stored in *Current Set*. Feedback is monitored by *Monitoring*, a supervisory process that clears *Current Set* in the event of negative feedback. When *Current Set* is empty, *Task Setting* is invoked. This process accesses *Working Memory* to gather details of previous unsuccessful sorting attempts (if any can be recalled) and generates a new potential sorting rule that has not been recently used. If there is more than one possible choice of rule consistent with current evidence, *Task Setting* chooses at random from the available choices.

The model is implemented in the C programming language. In order to ensure their independent nature, the supervisory sub-processes and Contention Scheduling are implemented as separate ‘threads’. Results are scored according to the criteria followed by Stuss et al. (2000).

Behaviour of the Model

The model’s behaviour is dependent on a number of parameters, which essentially determine the efficiency of processing of the various sub-processes. Table 1 provides a brief description of what these parameters represent and the range of values they can take. There are essentially two types of parameters: activation-related parameters (thresholds, activation persistence and activation boost parameters) and efficiency-related parameters. The latter specify the probability of a subsystem functioning. For instance, a value of 0.10 for Monitoring_{exogenous} specifies that monitoring is active roughly 10% of the time. The remaining 90% of the time, the process does not function.

All parameters have optimal or ideal values. Thus, when all efficiency parameters are set to 1.00, activation boost rates are set to the reciprocals of corresponding persistence rates (so that maintenance exactly counteracts decay), and thresholds are set to 0.5, the model sorts optimally, correctly sorting approximately 56 out of 64 cards, achieving 5 categories (i.e., correctly sorting 10 cards according to 5 different rules) and making errors only when it is attempting to discover a rule following negative feedback.

Modelling Control Performance

When neurologically healthy subjects attempt the WCST they generally do not perform at the optimal level. Thus the control subjects of Stuss et al. (2000) achieved on average 3.9 categories, made occasional perseverative errors, where they continued sorting by a rule even given negative feedback, and also occasionally produced set-loss errors, where they appeared to correctly infer the sorting rule, only to forget it even though the feedback was positive (see figure 3, right-most bars). In order to model control performance, normally distributed random noise was added to activation values of *Working Memory* and *Current Set* elements, the persistence of activation values was decreased, and the efficiency of supervisory processes was decreased. Systematic exploration yielded performance similar to controls when these parameters were set as follows: noise standard deviation = 0.05; Monitoring_{exogenous} = Monitoring_{endogenous} = Taskset_{none} = Taskset_{random} = 0.90; Memory_{persistence} = 0.70; and Attentiveness_{persistence} = 0.76. With these values, the model generates perseveration and set-loss errors at rates comparable to those of the control subjects of Stuss et al. (2000) – see figure 3. The values indicate that control performance can be modelled by introducing slight imperfections to the Supervisory System.

Modelling Frontal Dysfunction

Based on the arguments of Shallice et al. (2008), we associate *exogenous monitoring* (Monitoring_{exogenous}), *task setting* (Taskset_{none}) and *attentiveness* (Attentiveness_{persistence}) with right dorsolateral, left dorsolateral and inferior medial prefrontal patients respectively. Although *endogenous monitoring* (Monitoring_{endogenous}) and *task random setting* (Taskset_{random}) are important elements of monitoring and task setting processes, analysis of the model’s behaviour revealed that they do not contribute significantly to the dependent measures and hence they have been excluded from the analysis of frontal dysfunction. Moreover we do

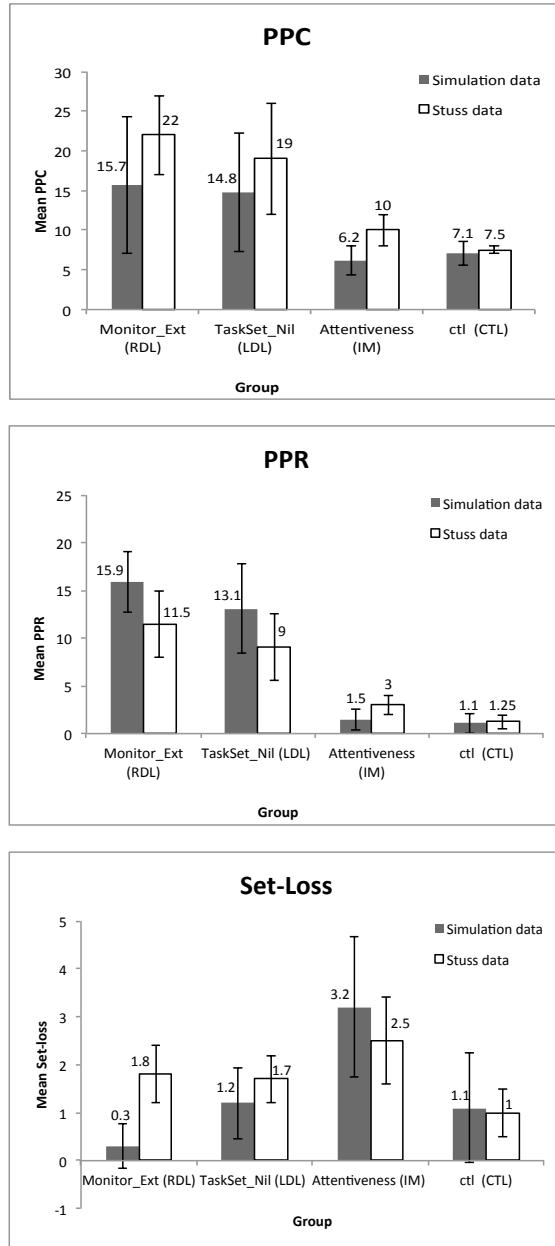


Figure 3: Model performance versus that of the patients of Stuss et al. (2000). Error bars represent one standard error from the mean.

not attempt to account for the behaviour of the superior medial prefrontal patients as the model does not have an explicit representation of the *energisation* process.

We adopt the methodological approach of modelling patient performance by reducing the efficiency of the process held to be impaired in the corresponding patient group. Specifically, we adjust the relevant parameter so that the model accurately captures the mean number of categories achieved by each set of patients in the Stuss et al. study (0.6 categories for RDL patients, 1.3 categories for LDL patients and 2.6 categories for IM patients), and then compare the model's behaviour on the three dependent measures described above (PPC, PPR and set-loss errors).

Thus an impairment level of 0.00 in Monitoring_{exogenous}, 0.10 in Taskset_{none}, and 0.74 in Attentiveness_{persistence} produced a mean category measure comparable to RDL, LDL and IM patients respectively. When setting these parameters to model the impairments of the three patient groups, all other parameters were fixed at the levels used to simulate control subjects. Simulation data on three dependent measures – PPC, PPR and set-loss errors – for each patient category obtained in this way and averaged over 10 runs of the model is shown in figure 3, plotted against the corresponding patient data published by Stuss et al.

General Discussion

As shown in figure 3, the model of WCST behaviour, embedded within the broader Supervisory System / Contention Scheduling architecture, is able to provide a good account of control subject behaviour across four dependent measures: categories obtained, PPC errors, PPR errors and set-loss errors. This provides support – albeit weak support – both for the Supervisory System / Contention Scheduling architecture and for the model of WCST within it. However, equally important for the current work is the behaviour of the model when damaged and its relation to that of neurological patients. When damaged in theoretically motivated ways, the model reproduces several key features of the behaviour of neurological patients. Most critically, an impairment of exogenous monitoring leads to elevated levels of PPC and PPR errors, as seen in right dorsolateral prefrontal patients. An impairment of task setting leads to a similar error profile, as seen in left dorsolateral prefrontal patients. Finally, an impairment of attentiveness leads to elevated set loss errors, as seen in patients with inferior medial prefrontal lesions. This provides further support for both the model and the association of these supervisory functions with the different regions of prefrontal cortex.

The results must be interpreted with caution, however. First, the model performs similarly with impairments to either exogenous monitoring or task setting. While this is consistent with patient behaviour, it supports an argument originally made by Stuss et al. (2000) that the erroneous behaviour of their right dorsolateral and left dorsolateral groups, whilst qualitatively and quantitatively similar, may in fact be due to different functional impairments. The model demonstrates that the WCST is unable to discriminate between these functional impairments (at least with respect to PPC and PPR errors), and that empirical studies of the two patient groups on other, more discriminating, tasks is necessary if one is to make the argument that the functions of (exogenous) monitoring and task setting are indeed supported by different regions of prefrontal cortex.

The reverse side of this argument, however, derives from the fact that inferior medial patients produce elevated numbers of set-loss errors but not of PPC or PPR errors. This pattern of behaviour is produced by an impairment to the effectiveness of the attentiveness sub-process. Thus the

model supports the treatment of 'attentiveness' as a functionally and structurally distinct sub-process, as well as the 'impaired attentiveness' account of inferior medial prefrontal patient performance.

A second caution regarding the results concerns the rate of set-loss errors in simulation of RDL patient performance, which is lower than that seen in patient behaviour. This is in part because the model must sort a minimum number of consecutively presented cards correctly (and hence demonstrate that it is following a rule) before an error can be counted as a set-loss error. With severe impairments in the model, this is rare. Hence the opportunity for set loss errors is rare. We have simulated the RDL patient group by setting $Monitoring_{exogenous}$ to 0.00 in order to match performance on the number of categories correctly sorted. Perhaps this level of impairment is too severe. This is an issue to be addressed in future work.

The issue of severity relates to the methodology employed in simulating patient behaviour. Patient performance was modelled by choosing one parameter value to match the number of categories achieved by the model to that of the relevant patient group. This does not take account of the heterogeneity of each patient group – not all patients were equally severely impaired – and a more appropriate methodology would be one that attempted to match the varying severity of individual patients, rather than of each group as a whole. This is an issue for further research, though in the absence of individual subject data, a plausible strategy may be to sample different levels of severity, as used by Cooper et al. (2005) in modelling the action errors of neurological patients.

Two more general questions concern the nature of supervisory processes and the Contention Scheduling / Supervisory System architecture within which the WCST model is embedded. Considering first the architectural issue, the model demonstrates that the functional decomposition of Contention Scheduling and the Supervisory System is able to support behaviour on a complex task, and so the Contention Scheduling / Supervisory System architecture provides a viable alternative to production system architectures such as ACT-R, Soar and EPIC. At the same time, the architecture remains relatively underspecified and substantial elaboration of the architecture and its subcomponents (e.g., through application to other tasks) is required before it can be fully compared with these alternatives.

With regard to the nature of supervisory processes, many theorists appear to assume, at least implicitly, a distinction between supervisory and non-supervisory processes. There is, however, some debate about whether the supervisory system is most appropriately viewed as a unitary system (e.g., Duncan, 2010) or as functionally heterogeneous (e.g., Stuss, & Benson, 1986; Shallice, & Cooper, 2011), and whether prefrontal cortex is better viewed as functionally hierarchical (e.g., Badre, 2008) or functionally heterarchical (Shallice & Burgess, 1996; Shallice et al., 2008). The model described in this paper substantiates the theoretical stand of

Shallice and colleagues, but again further work is required. A possible extension to the present model is to generalise it to other executive tasks such as Tower of Hanoi, Tower of London, Stroop test etc. Applying the architecture to other executive tasks will allow for better validation of the theoretical hypotheses, which are not adequately and independently assessed by WCST.

References

- Anderson, J. R. (2007). *How Can the Human mind Occur in the Physical Universe?* Oxford University Press, Oxford, UK.
- Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5), 193-200.
- Cooper, R.P., Schwartz, M.F., Shallice, T. & Yule, P. (2005). The simulation of action disorganisation in complex activities of daily living. *Cognitive Neuropsychology*, 22, 959–1004.
- Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Science*, 14, 172-179.
- Kieras, D. E. & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12, 391-438.
- Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology*, 9, 90-100.
- Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.
- Norman, D. A. & Shallice, T. (1986). Attention to action: willed and automatic control of behaviour. In R. Davidson, G. Schwartz & D. Shapiro (eds.) *Consciousness and Self Regulation, Volume 4*. NY: Plenum.
- Shallice, T., & Burgess, P.W. (1996). The domain of supervisory processes and temporal organisation of behaviour. *Philosophical Transactions of the Royal Society of London*, B351, 1405–1412.
- Shallice, T., & Cooper, R.P. (2011). *The Organisation of Mind*. Oxford: Oxford University Press.
- Shallice, T., Stuss, D. T., Picton, T. W., Alexander, M. P., & Gillingham, S. (2008). Mapping task switching in frontal cortex through neuropsychological group studies. *Frontiers in Neuroscience*, 2, 79-85.
- Sood, M. R. (2012). *Executive Function Deficits of Neurological Patients with Frontal lobe injury: Analysis using Computational Modelling*. Unpublished master's thesis. Birkbeck, University of London, United Kingdom.
- Stuss, D. T., & Benson, D. F. (1986). *The Frontal Lobes*. New York: Raven Press
- Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., Murphy, K.J. & Izukawa, D. (2000). Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, 38, 388–402.