BIROn - Birkbeck Institutional Research Online

Cooper, Richard P. (2016) Executive functions and the generation of "Random" sequential responses: a computational account. Journal of Mathematical Psychology 73 (1), pp. 153-168. ISSN 0022-2496.

Downloaded from: https://eprints.bbk.ac.uk/id/eprint/15406/

Usage Guidelines:
Please refer to usage guidelines at https://eprints.bbk.ac.uk/policies.html or alternatively contact lib-eprints@bbk.ac.uk.

To appear in the *Journal of Mathematical Psychology*
Final draft. June 2016

Executive Functions and the Generation of "Random" Sequential Responses: A Computational Account

Author: Richard P. Cooper

Affiliation: Centre for Cognition, Computation and Modelling
Department of Psychological Sciences
Birkbeck, University of London

Telephone: +44 20 7631 6211

Email: R.Cooper@bbk.ac.uk

Street Address: Department of Psychological Sciences,
Birkbeck, University of London,
Malet Street, London, WC1E 7HX
United Kingdom

Executive Functions and the Generation of "Random" Sequential Responses: A Computational Account

Richard P. Cooper

Centre for Cognition, Computation and Modelling
Department of Psychological Sciences
Birkbeck, University of London

Abstract: When asked to generate sequences of random responses, people exhibit strong and reliable biases in their behaviour. The origins of these biases have been linked to the operation of so-called executive functions through empirical studies varying, e.g., rate of production, modality of response, and (in dual task conditions) secondary task. We present a computational process model of random generation that accounts for a broad range of these empirical effects. The model, which operationalises a previous verbal account of random generation, is grounded in both the cognitive architectures and the executive functions literatures. As such, it instantiates a hypothesis concerning the interaction of multiple distinct executive functions in the generation of complex behaviour. In particular, it is argued on the basis of simulations of empirical findings that three cognitive factors play separable roles in random generation behaviour: cognitive load, which when high exacerbates underlying biases in a generation stage, monitoring, which when impaired results in greater inequality of response usage, and set-shifting, which when impaired results in less frequent switching between response schemas.

Highlights:

We present an information-processing model of random generation behaviour

The model is developed within a more general cognitive architecture

The architecture makes explicit the role of executive functions

The model reproduces random generation performance in single and dual task conditions

Dual task performance is accounted for via interference to executive control

Keywords:

Executive function; information processing model; random generation; set shifting; monitoring; response inhibition.

1 Introduction

One goal of much empirical work over the past two decades, in both cognitive psychology and cognitive neuroscience, has been to characterise the operation of the "central executive" in terms of separable functions or processes such as task-setting, working memory maintenance, response selection, monitoring and response inhibition (for reviews, see for example: Aron, Robbins & Poldrack, 2014; Cooper, 2010; Nee et al., 2013; Vandierendonk, Liefooghe & Verbruggen, 2010). There is now considerable empirical support for many such functions, and several computational accounts of specific executive functions have been developed (e.g., task switching: Gilbert & Shallice, 2002; Brown, Reynolds & Braver, 2007; Altmann & Gray, 2008; Herd et al., 2014; memory maintenance and updating: Ashby, Ell, Valentin & Casale, 2005; O'Reilly & Frank, 2006; response inhibition: Band et al., 2003;

Boucher, Palmeri, Logan & Schall, 2007; Wiecki & Frank, 2013). A limitation of much of this work – though one that is reasonable given the developing nature of the field – is that it generally considers the operation of individual executive functions in isolation on relatively simple executive function tasks. For example, the Boucher et al. (2007) model focuses on a version of the stop-signal task, in which a subject must, on a small proportion of trials normally signalled by a tone, withhold the production of an otherwise routine response such as a button press. Working at a similar level of complexity, Gilbert and Shallice (2002) focus specifically on the process of task set switching involved in switching from word reading to colour naming (and vice versa) in the Stroop task. While this work has advanced our understanding of basic executive functions, behaviour on tasks of greater complexity is likely to require multiple executive functions working in concert in order to appropriately regulate behaviour.

A landmark empirical study specifically aimed at understanding the role of executive functions in more complex tasks is the individual differences study of Miyake et al. (2000). These authors considered the correlational structure of the performance of over 130 individuals on 14 tasks, comprising 9 relatively simple tasks and 5 more complex tasks. The simple tasks included three held primarily to tap the putative executive function of set-shifting, three to tap memory updating and monitoring, and three to tap response inhibition. Subject performance within each subset of three simple tasks was found to correlate more highly than between subsets, as supported by Confirmatory Factor Analysis (CFA). The three factors obtained from this analysis were then used via Structural Equation Modelling (SEM) to determine the role of the three underlying constructs in the more complex tasks. Thus, on the basis of this it was argued that, for example, in the Wisconsin Card Sorting Test the set-shifting function (and this function alone) was critical in limiting the production of perseverative errors. This was in contrast to an alternative hypothesis, namely that such errors arose from a failure in response inhibition.

One of the five complex executive tasks used by Miyake et al. (2000) was random number generation. In random generation tasks, which are widely used in executive function research (e.g., Baddeley, Emslie, Kolodny & Duncan, 1998; Cooper, Wutke & Davelaar, 2012; Jahanshahi et al., 1998, 2000, 2006; Peters, Giesbrecht, Jelicic & Merckelbach, 2007; Proios, Asaridou & Brugger, 2008; Towse, 1998; Towse et al., 2016), subjects are asked to produce a sequence of "random" responses. More precisely, subjects are required to produce responses where each successive response is independent of his or her previous responses, or equivalently, where each successive response cannot be predicted with greater than chance accuracy from the subject's previous responses. Random generation tasks typically yield multiple dissociable dependent measures (as discussed in detail below; see also Towse & Neil, 1998), and Miyake et al. argue on the basis of their SEM analysis that different dependent measures reflect the efficacy of different executive functions. This paper aims to evaluate the account of executive involvement in random generation performance given by Miyake et al. and, more generally, to explore the interaction of executive functions on a task with multiple dependent measures. We present a computational account of random generation derived from the verbal model of Baddeley et al. (1998) and grounded in a cognitive architecture in which executive functions have explicit roles. The model, which is a development of that of Sexton and Cooper (2014), is evaluated both against the results of Miyake et al. and against results from a dual task study of Cooper et al. (2012) where the primary task was random generation.

The remainder of this paper is structured as follows. We begin by providing an overview of previous findings from research on random generation tasks together with a re-analysis of the dual-task interference data from Cooper et al. (2012). We then present a computational model of performance on a random generation task. The model builds upon

the verbal model of Baddeley et al. (1998), embedding it within a cognitive architecture based on the Contention Scheduling / Supervisory System approach of Norman and Shallice (1986). Critically, the Supervisory System part of the model draws also on the executive functions investigated by Miyake et al. (2000) and the decomposition of supervisory processes outlined by Shallice and Burgess (1996). Subsequent sections report simulation studies that demonstrate that the model can capture both frequently reported biases in random generation and the interference patterns arising from concurrent performance of different secondary tasks. The general discussion focuses on two broad sets of issues: the role of key parameters of the model and their relation to the efficacy of executive functions; and the model's architecture as an elaboration of the verbal theories on which it is based, with a specific focus on the relationship between the model's architecture and other established cognitive architectures (including ACT-R: Anderson, 1993, 2007, and Soar: Newell, 1990; Laird, 2012).

2 Random Generation

2.1 Dependent Measures and Standard Effects

Random generation tasks have a long history within information processing psychology research (see Tunes, 1964, and Wagenaar, 1972, for early reviews). Behaviour is typically assessed through measures of randomness calculated from the sequence of responses produced by the subject. The degree of randomness of a sequence cannot be characterised with a single measure and many different measures have been considered. Thus, if each successive response in a sequence is equally likely but independent of the previous responses then one would expect, over the long run, that each response would occur equally often. The degree of response equality is frequently quantified in information-theoretic terms by the redundancy, or *R*, score:

$$R = 100 \times \left(1 - \frac{\log_2 n - \frac{1}{n} \sum_i n_i . \log_2 n_i}{\log_2 a} \right) \tag{1}$$

where n is the number of responses in the sequence,
a is the size of the response set,
and n_i is the number of times response i is produced.

The redundancy score, which is a linear transformation of the Shannon entropy (Shannon, 1948) of the multiset of elements in sequence, ranges from 0 to 100 and is 0 when each response is equally frequent and 100 when all responses are identical (i.e., when a single response is repeated).

A highly unpredictable sequence will have a low redundancy score, but cycling through all possible responses will similarly produce a low redundancy score. It is therefore necessary to also consider the relative frequency of response pairs (i.e., of bigrams). Again, in an ideal random sequence of infinite length each bigram should be equally frequent. Bigram equality may be quantified by calculating a redundancy score for bigrams (rather than individual responses) using equation 1. An alternative approach is the RNG score of Evans (1978):

$$RNG = \frac{\sum_{ij} n_{ij} . \log_2 n_{ij}}{\sum_i n_i . \log_2 n_i} \tag{2}$$

where n_i is the frequency of response i
and n_{ij} is the frequency of response i followed by response j.

The RNG score scales the entropy of bigrams by the entropy of elements. It attains a maximum of 1 when an element is perfectly predicted by its predecessor (i.e., bigrams probabilities are zero or one). Lower values reflect greater equality of bigram usage. Note, however, that even if a sequence has perfect equality of bigram usage, it may still be predictable at the level of trigram (or higher) statistics. Further measures of randomness are therefore required. Indeed, Towse and Neil (1998) report 11 measures that have been used to quantify randomness, while Towse and Valentine (1997) consider 16.

The extensive empirical work on random generation has yielded several robust behavioural effects and substantial convergence on a verbal account of the underlying cognitive processes. With regard to the former, five findings in particular have been replicated by a number of studies:

1. Generation is more stereotyped when the response set is externally represented (e.g., subjects select from a visual array) than when the subject must hold the response set in mind (e.g., subjects select according to a criterion such as digits ranging from 1 to 10). See for example Baddeley et al. (1998) and Towse (1998), and see Wagenaar (1972) for a review of earlier studies.
2. Generation is affected by biases inherent in the response set or response modality (e.g., when responses are digits counting up/down is more likely than chance, when responses are letters associated letter pairs are more likely than chance, when responses are generated with keypads homologous finger selection on alternating hands is more likely than chance). See Baddeley (1966), Wiegersma (1984), Baddeley et al. (1998), Towse (1998) and Cooper et al. (2012).
3. Subjects produce far fewer repeat responses (e.g., “3” followed by “3” when generating digits) than would be expected if successive responses were independent. See, for example, Baddeley et al. (1998), Towse (1998), Cooper et al. (2012) and Towse et al. (2016).
4. Generation is more stereotyped when the response set is large (e.g., 15 items or more) than when it is small (e.g., 10 items or fewer). See Towse (1998) and Towse & Vallentine, (1997), and for a review of earlier work see Tunes (1964) and Brugger (1997).
5. When response generation is paced, responses are more stereotyped when response rate is fast (e.g., less than one per second) than slow (e.g., one per four seconds). See Baddeley (1966), Towse (1998) and Jahanshahi et al. (2000, 2006), and for a review of earlier work see Tunes (1964), Wagenaar (1972) and Brugger (1997).

Findings 2 and 3 relate specifically to biases in bigram production. In the other cases, R and RNG (where both are reported) generally behave analogously, with both showing a tendency towards increased stereotypy or predictability in the same conditions. However, R and RNG have also been shown to dissociate. For example, Towse and Valentine (1997) scored the randomly generated sequences of their subjects on 16 measures of randomness. They found considerable individual differences on these measures. Factor analysis suggested four dissociable factors underlying performance, which they labelled *equality of response usage*, *prepotent associates*, *short repetitions*, and *long repetitions*. The first of these correlates most strongly with R, while the second correlates more strongly with RNG and bigram measures (such as the adjacency index – an index of the relatively frequency of adjacent response pairs).

2.2 The Role of Executive Functions in Random Generation

Building on the factors of Towse and Valentine (1997), the SEM analysis of individual differences by Miyake et al. (2000) found that *equality of response usage* was best accounted for by a model with one path from the executive construct of “memory monitoring and

updating" while *prepotent associates* was best accounted for by a model with one path from the executive construct of "response inhibition" (in both cases in contrast to models with paths from constructs related to task set shifting or memory updating and monitoring, or all three executive function constructs).

Cooper et al. (2012) argued that a potential limitation of the Miyake et al. (2000) study was its reliance on correlational rather than experimental methods. Miyake et al.'s conclusions from SEM analyses, in particular, depend on selecting the best structural model from a set of competing models, all of which have qualitatively similar fits. Cooper et al.'s approach was instead to consider interference on random generation due to different secondary tasks, where the secondary tasks were designed to tap different executive functions. Thus, when subjects attempted to generate sequences of random responses while concurrently completing the well-known 2-back task – a task held primarily to tap the putative executive function of memory monitoring and updating – the sequences so generated were far from random. In comparison to a baseline condition, the memory monitoring and updating task had a substantial inflationary effect on both R (reflecting decreased equality of response usage) and RNG (reflecting increased frequency of prepotent associates). A concurrent task designed to tap set-shifting – the digit-switching task of Monsell (2003) – had a similar substantial inflationary effect on R but a more modest effect on RNG. Finally, a concurrent task designed to tap response inhibition – a version of the popular go-no go task – had a modest inflationary effect on both R and RNG. Cooper et al. argued that the various patterns of interference in random generation, and in particular the dissociation between dependent measures occurring with the set-shifting task, reflected interference to different executive functions involved in the random generation task.

This apparent fractionation of executive function is important because it addresses a criticism sometimes levelled at theoretical accounts of executive function, namely that the oft-posited central executive is homuncular (e.g., Dennett, 1998). The existence of dissociable factors suggests dissociable processes underlying those factors, and hence that the central executive is not an atomic, non-decomposable, entity. Verbal accounts of the cognitive processes underlying random generation, as well as neuroscientific evidence, also support this position. Thus, Baddeley (1996; see also Baddeley et al., 1998) suggested that random generation involves a series of steps. Subjects, he suggests, use response schemas (like counting up by two, or down by one) to generate a putative response based on the previous response. This putative response is evaluated in the context of recent responses to determine whether it is "random". If the putative response passes this evaluation, it is produced. Otherwise (and if time permits) the subject inhibits production of the putative response, switches to a new response schema, and generates a new putative response. While Baddeley did not elaborate this verbal account into a complete computational process model of random generation, it clearly draws on several distinct processes: one process to generate a putative response based on a previous response and a response schema; a monitoring process to evaluate whether the putative response is sufficiently random; and, if it is not, processes to suppress production of the putative response (i.e., a response inhibition process) and to generate or switch to an alternative response schema (i.e., a set shifting process). The model is therefore compatible with the Miyake et al. (2000) decomposition of executive processes.

Evidence from neuroscience also supports this verbal account. Jahanshahi et al. (1998) applied Transcranial Magnetic Stimulation (TMS) over various regions of cortex while subjects performed a verbal random number generation task. TMS applied over left dorsolateral prefrontal cortex (DLPFC) resulted in a significant increase in the habitual response of counting up or down (e.g., 3 – 4 – 5, or 8 – 7 – 6) with an associated decrease in counting up or down by two (e.g., 3 – 5 – 7, or 8 – 6 – 4). This pattern was not observed with TMS over right DLPFC or medial frontal cortex. Jahanshahi et al. interpret their results in

Table 1: Expected values of measures of randomness and their values from the dual-task study of Cooper et al. (2012). R is response redundancy, as calculated from Equation 1. RNG is equality of bigram usage – see Equation 2. RR is the proportion of repeat responses in the sequence. AR is the proportion of response pairs that are adjacent, while OR is the proportion of response pairs that are diametrically opposite on the clock face layout. TPI, the turning point index, reflects the tendency to switch between clockwise and counter clockwise selection of responses. Its value is here normalised to 1.0 for an unbiased sequence. Values shown in the Expected Value column are means computed from 5,000,000 computer-generated pseudo-random sequences. Values shown in the Baseline Condition column are calculated from responses from human subjects performing the random generation task in isolation. Values shown in the Secondary Task columns are derived from dual-task performance of the same subjects with the given secondary task.

Dependent Measure	Expected Value (Random sequence)	Baseline Condition	Secondary Task		
			Digit-Switching	2-Back	Go / No-Go
R	1.995	0.962	2.048	1.979	1.196
RNG	0.242	0.300	0.410	0.461	0.388
RR	0.100	0.014	0.004	0.002	0.005
AR	0.200	0.259	0.328	0.424	0.334
OR	0.100	0.131	0.136	0.097	0.130
TPI	1.000	0.733	0.604	0.446	0.582

terms of a model where left DLPFC provides top-down control that inhibits habitual schemas such as counting up or down by 1. A follow-up imaging study using PET was consistent with this interpretation (Jahanshahi et al., 2000).

2.3 A Re-examination of the Data of Cooper et al. (2012)

For the purposes of being concrete, this article considers in depth the random generation data from the dual-task study of Cooper et al. (2012). That study asked subjects to generate sequences of 100 responses selected using a computer mouse from a clock-face type display with 10 response options arranged around a central cursor. Selection was self-paced and the mouse automatically returned to the central cursor after each item was selected. The procedure was performed once with no secondary task and then three further times – once with each of three different secondary tasks (with secondary task order fully counterbalanced). Full details of the procedure are given in the original paper (Cooper et al., 2012). Key measures of randomness are reproduced in Table 1, where "expected values" are calculated from computer-generated pseudo-random sequences of 10 items and length 100.

Several aspects of the dataset are of note. First, while the expected value of R (cf. Equation 1) for an *infinite* sequence is zero, its expected value for *finite* sequences is greater than zero.[1] Note also that the observed value of R in the baseline condition was less than the

[1] To understand why, consider a sequence of 100 items with a response set of size 10 (as in the Cooper et al., 2012 study). R will have a value of zero if and only if the generated sequence contains exactly 10 instances of each response. In a random sequence this is highly unlikely. Suppose an agent is on target to produce such a sequence, and so after 99 responses all but one item has been produced 10 items, with the remaining item (say "A") having been produced just 9 times. If responses are independent then there is clearly only a one in ten

expected value (and significantly so: see Cooper et al., 2012). This indicates that (under baseline conditions) human subjects were better able to equate response usage than would be expected by chance. This was not the case, however, under two of the secondary task conditions, where equality of response usage increased to the expected value.

The RNG statistic (cf. Equation 2) behaved differently. Like R, its expected value for a finite sequence is greater than zero, but its observed value under baseline conditions was greater than the expected value (indicating significant bias in bigram usage), and increased with concurrent performance of a secondary task. Of note, the RNG statistic was greatest when random generation was paired with a demanding memory task (the 2-back task).

RR, AR and OR all reflect specific bigram biases and can be understood as breaking down the RNG score into separate components. As in most studies of random generation, repeat responses were rare, with only 0.014 responses being repeats in the baseline condition. However, repeat responses were significantly less frequent in the secondary task conditions, suggesting that repeat responses are not suppressed by subjects in the baseline condition due to a biased concept of randomness, but are in some sense deliberately generated in a way that is impaired by secondary task performance (where repeat responses were almost never produced). Under baseline or control conditions subjects produced more adjacent response pairs than would be expected by chance (i.e., AR > 0.20), and this tendency increased under dual-task conditions – particular with the demanding memory task. Opposite response pairs were also produced more frequently in the baseline condition than would be expected by chance (i.e., OR > 0.10), but this measure was not affected by concurrent performance of two different secondary tasks, and *decreased* with concurrent performance of the demanding memory task – possibly as a consequence of the increase in adjacent response pairs in this condition.

The inter-relations between RR, AR, OR and RNG under the four experimental conditions are clarified by Figure 1, which shows the relative frequency of each type of bigram – repeats, rotations one step clockwise, two steps clockwise, etc. – under the four conditions. The figure illustrates that in the baseline condition there is a general clockwise bias, superimposed upon a bias to adjacent and opposite responses. These biases are magnified in the dual-task conditions, with an increased tendency to adjacent responses coupled with a decreased tendency to repeat responses.

The final statistic, the turning point index (TPI), reflects the tendency to switch between clockwise and counter-clockwise selection on the display. Its value is scaled so that it is 1.0 for an unbiased sequence, with higher values reflecting more frequent switching than chance and low values reflecting a tendency to persist in one direction longer than expected. Under baseline conditions subjects exhibited this tendency (to persist in one direction) – a tendency that was exacerbated under dual-task conditions (with the demanding memory task again having the greatest effect).

The dataset is suggestive of several features of the processes underlying random generation:

a) In order to achieve lower R scores than chance, subjects must either adopt a strategy that ensures good equality of response usage as a side effect (e.g., cycling through responses), or maintain some record of their responses that allows them to avoid choosing any one response too often. There is no evidence to suggest the former.[2]

---

chance, even from this unlikely position, of producing the critical item ("A") on the last response. Thus the probability of generating exactly 10 instances of each response is low. The expected value of R decreases to zero as sequence length increases to infinity.

[2] If subjects were adopting a strategy that minimised R in the baseline condition, then this should be apparent in other measures of randomness, and those other measures should be less compromised in the secondary task conditions, where R is closer to the expected value. None of the other measures show this pattern.

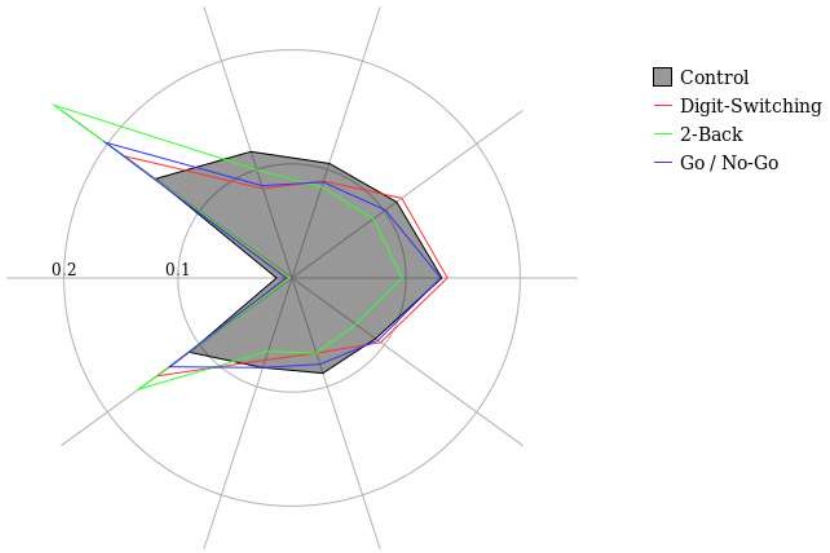

Figure 1: Relative bigram frequencies in the study of Cooper et al. (2012). Radial lines represent the axes corresponding to bigrams of each type, with repeat responses shown on the horizontal axis to the left of center, bigrams whose second constituent is one unit clockwise from the first shown on the first radial line clockwise from the horizontal, bigrams whose second constituent is two units clockwise from the first shown on the second radial line clockwise from the horizontal, and so forth. Each polygon represents one experimental condition, with the baseline condition shown by the grey filled polygon.

Assuming the latter, this record appears to be compromised (or ignored) in two of the dual-task conditions (resulting, paradoxically, in R scores closer to that expected by chance);

b) The bigram data suggest an underlying bias towards adjacent and opposite responses but away from repeat responses. This bias is magnified in dual task conditions, particularly when the secondary task is the 2-back task; and

c) These two features dissociate, with R resorting to its expected value for two of the secondary task conditions, but bigram statistics being most severely distorted by just one of those conditions.

These features are at least superficially consistent with other studies that suggest dissociable processes underlie random generation, such as those of Towse and Valentine (1997) and Miyake et al. (2000) described above.

3 A Model of Random Generation

The account of random generation offered by Baddeley (1996) and described above is limited in two ways. First, it is specified verbally and therefore open to (mis-)interpretation. Second, it is not specified in the terms of the functional components required to fully operationalise the model. This section presents a process model of random generation based on the verbal account that is a) computationally fully specified, and b) anchored in a more general view of the functional organization of mind.

3.1 Theoretical Background

The steps of Baddeley et al.'s account of random generation as described above suggest separable processes for the generation of schema-driven behaviour, the continuous monitoring of behaviour, and the selection or switching of control schemas when appropriate. As noted by Baddeley (1996), such a decomposition of cognitive processing is consistent with the Contention Scheduling / Supervisory System theory of Norman and Shallice (1986; see also Shallice & Burgess, 1996). Within this theory, all behaviour is controlled by schemas

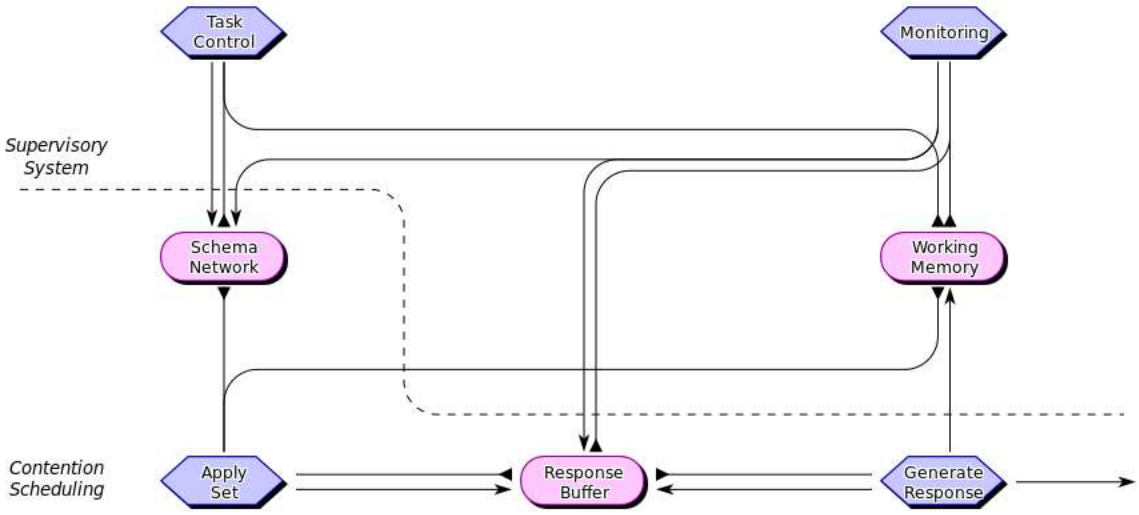

Figure 2: The functional components of the model of random generation. Rounded rectangles represent buffers, which support the temporary storage of information. Hexagonal boxes represent processes, which operate on the contents of those buffers. Flat-ended arrows represent read operations. Sharp-ended arrows represent write operations. All components operate in parallel.

– partially ordered sets of lower-level behaviours. Schema selection is determined by an activation-based competitive process (contention scheduling). This process operates autonomously in routine situations, with schemas receiving excitation or inhibition from an internal model of the environment and from higher level schemas. (See Cooper & Shallice, 2000, for an implementation.) In non-routine situations the operation of contention scheduling is modulated by a second system – the supervisory system – which is held to operate indirectly on behaviour by selectively exciting or inhibiting schema representations within the contention scheduling system.

3.2 Verbal Description of the Model and its Operation

Figure 2 depicts the functional components of the model of random generation.[3] The figure adopts a formal graphical language for describing functional decomposition within information processing models in which hexagonal boxes represent processes that transform information, rounded rectangles represent buffers that store information, pointed arrows represent information being sent or written to one component by a process, and flat-ended arrows represent information being read from a buffer by a process (see Cooper & Fox, 1998). All components operate in parallel. The upper portion of the diagram, which operationalises a simplification of the supervisory system, comprises two processes and one buffer. The lower portion, which operationalises the contention scheduling system, comprises two buffers and two processes.

Consider first the components and operation of contention scheduling. *Schema Network* contains the set of possible schemas that might be used to generate a response, at most of one

[3] The model and functional architecture presented here is a simplification of a previous model of random generation which was focussed on understanding the effect of rate on random generation performance (Sexton & Cooper, 2014). That model was specified within essentially the same architecture as shown in figure 2. However, the effects of varying the efficacy of the various executive functions were not explored in that model. Equally, that work did not attempt to address dual task interference. Full details of all components of the current model are given in Appendix A. The complete source code (in C) is available at http://www.ccnl.bbk.ac.uk/models/rng_2016.tgz.

of which is active (i.e., controlling behaviour) at any time. *Apply Set* implements the assumption that successive random responses are produced by the application of the active schema to previous responses. Thus, *Apply Set* proposes a possible response by consulting *Working Memory* (for the previously generated item, e.g., "3") and *Schema Network* (for the active schema, e.g., "+4"). The putative response (e.g., "7") is stored in *Response Buffer*, which is a temporary store that allows a putative response to be vetted before it is actually produced. Production of the response is the job of *Generate Response*. In paced random generation *Generate Response* will be triggered by the pacing signal (e.g., a metronome). In self-paced random generation it is assumed that *Generate Response* normally waits for supervisory processes (specifically *Monitoring*) to check that a proposed response is sufficiently random before outputting it. On production of a response, *Generate Response* also copies the response to *Working Memory*, a component of the supervisory system that stores recent responses and is used both to provide a seed for *Apply Set* and to ensure (via *Monitoring*) that successive responses are sufficiently random.

The active schema in *Schema Network* is set by *Task Control*, a supervisory system process that is invoked automatically when there is no active schema within *Schema Network*. This may occur if the active schema decays through inadequate maintenance, or if *Monitoring* rejects the current schema. *Monitoring* is a supervisory process that examines a putative response (as stored in *Response Buffer*), decides if it is sufficiently random (given the history of recent responses stored in *Working Memory*) and, if not, inhibits production of the response (by clearing *Response Buffer*) and switches the current schema (by deactivating the active schema in *Schema Network*).

The main work of the model is done by the two supervisory processes. *Monitoring* must decide, based on the contents of *Working Memory* and *Response Buffer* whether a candidate response is sufficiently random. *Working Memory* includes a (possibly imperfect) record of recently generated responses, while *Response Buffer* contains the candidate for the next response. Conceivably there are individual differences in what subjects consider to be "sufficiently random", but we assume a minimal condition, namely that the candidate response is not a recent response (i.e., is not in *Working Memory*). *Task Control*, on the other hand, must propose a schema when none is available. As a working hypothesis, we assume that *Task Control* involves selecting at random from the available schemas, subject to biases imposed by the response modality. Thus, if responses are numbers given verbally, then *Task Control* will show a bias towards selecting schemas such as +1 or –1, and away from, e.g., +7. In contrast, if responses are given by selecting from a clock-face display, as in the experiments of Cooper et al. (2012), then *Task Control* will show a bias towards adjacent and opposite schemas, and away from schemas which select, say, three positions clockwise from the previous response.

As noted earlier, the empirical data indicate that response-modality-induced biases in schema selection are more pronounced under fast paced conditions (Baddeley, 1966; Towse, 1998), when TMS is applied over left DLPFC (Jahanshahi et al., 1998) and under dual-task conditions (Cooper et al., 2012). Given this, we assume that each schema S_i has a weight w_i, and the probability of selection of schema S_i is given by the Boltzmann equation:

$$p(S_i) = \frac{e^{w_i/\tau}}{\sum_j e^{w_j/\tau}} \tag{3}$$

where τ is the temperature parameter. At very high temperatures selection is equi-probable. At low temperatures biases inherent in response schemas dominate. It is further assumed that weights are fixed (within individuals), but that temperature reflects the availability of

cognitive resources and varies with task conditions. Full details of the specific schemas and schema weights assumed in the simulations reported here are given in Appendix B.

One further requirement of *Task Control* is that it should frequently switch the schema guiding the proposal of responses, regardless of whether *Monitoring* considers a putative response to be insufficiently random. Whatever schema is active, generating a sequence of responses with that schema will ensure that the responses are not random. To avoid this, *Task Control* periodically deactivates the active schema. Given the processes already described, an alternative schema will then be activated (with the alternative being chosen according to equation 3).

3.3 Parameters of the Model and their Putative Relation to Executive Functions

In addition to the weights of each schema, the model has five key parameters. First, responses are recorded in *Working Memory*, but it is assumed that decay operates on the contents of *Working Memory*. Second, recording of responses in *Working Memory* may be prone to failure. These lead to two distinct parameters: *WM Decay Time* (the maximum number of processing cycles an element can remain in *Working Memory*, with probabilistic deletion prior to this) and *WM Update Efficiency* (the probability that updating is successful). Third, according to the competence model, when a putative response is placed in *Response Buffer* it is checked by *Monitoring* to ensure that it is sufficiently random (subject to the model's conception of randomness). Again, at the performance level it is conceivable that this process does not operate on every putative response, and this leads to a third parameter: *Monitoring Efficiency* (the probability that monitoring rules are invoked). Finally, two parameters govern the operation of *Task Control*: *Switch Rate* (the probability of switching schemas after each response), and *Temperature* (the value of τ in equation 3).

The approach to parameterisation aims to clarify potential sources of variation within the model. This results in many, rather than few, parameters. While it might be possible to consolidate some parameters by linking them (e.g., through a single efficiency parameter), this would fail to acknowledge the functional modularity assumed within the underlying architecture and artificially constrain the resultant model. Our approach is rather to be over-zealous, if anything, in the identification of parameters, but to guard against overfitting by then consider the effects (including interaction effects) of variation of each parameter, to fix parameters where they do not affect the model's behaviour (e.g., because variation in one parameter can be countered by variation of another parameter), and to compare the behaviour of the fully parameterised model with that of restricted versions (e.g., in which parameters are set to optimal values or where parameters are linked). Supporting simulations demonstrating the effects of varying individual parameters are reported in the Supplementary Materials.

4 Simulation 1: The Generation of Random Response Sequences

The purpose of simulation 1 was to demonstrate that, with appropriate parameter settings, the model is able to capture the general patterns in performance observed in human subjects on a random generation task (i.e., the ballpark values for a range of measures of randomness). The target data were those from the baseline condition of experiment 1 of Cooper et al. (2012), as described above. These data were used as a target because a) in contrast to many studies (e.g., those of Baddeley et al., 1998 or Miyake et al., 2000) they include multiple dependent measures and b) the raw data are openly accessible[4], allowing additional analyses to be performed for the current work and supporting further analysis by researchers not involved in the original research.

[4] The data are available at http://www.ccnl.bbk.ac.uk/data/qjep_2012_ex1.tgz

4.1 Method

Preliminary simulations with the model revealed that human-like behaviour appeared to occur when:

a) *WM Decay Time* was approximately 15;
b) schema selection was biased to favour adjacent and opposite responses over other responses, and clockwise responses over count-clockwise responses;
c) schema biases were modulated on a subject-by-subject basis by a moderate level of noise to account for individual differences in those biases (see Appendix B for further discussion of schema biases, schema bias noise, and the precise values of the biases for each schema); and
d) other parameters (*WM Update Efficiency*, *Monitoring Efficiency* and *Switch Rate*) were set to maximal values (1.00 in each case).

Additional simulation work found that the effect of increasing *WM Decay Time* on each of the dependent measures discussed above (R, RNG, RR, AR, OR and TPI) could be countered by decreasing *Monitoring Efficiency*. In order to avoid artefacts due to bounds on parameters the simulations reported in this section therefore assume *WM Decay Time* of 30 and *Monitoring Efficiency* of 0.65. τ was fixed at 1.00 and weights for each schema were set by hand to generate sequences that appeared, with the naked eye, to yield values of randomness similar to those observed by Cooper et al. (2012). Thirty-six sequences each of 100 simulated responses, corresponding to thirty-six virtual subjects, were then generated by the model. (Note that at this stage no formal method was used to maximise the fit of the model to the human data.) Each simulated sequence was scored on the measures described above, namely R, RNG, RR, AR, OR and TPI.

For comparison, thirty-six pseudo-random sequences of 100 responses were also generated using an unbiased random number generator and scored in the same way in order to estimate chance values (and standard deviations) of the various measures. The procedure therefore yielded three groups of comparable scores, with each derived from 36 (real or virtual) subjects, corresponding to measures of randomness for: the human subject data of Cooper et al. (2012); chance sequences of the same length; and simulated sequences of the same length.

4.2 Results

Means and standard deviations of all dependent measures for the three groups – human subjects of Cooper et al. (2012), chance, and the simulated subjects – are shown in table 2. Inspection of the data suggests that the human data show biases on most if not all measures in comparison to chance. These biases appear to be replicated in the sequences produced by the model. Pairwise between-subjects *t*-tests revealed that for all dependent measures and in all but one case: a) baseline subject values differed significantly from chance values; b) simulated values differed significantly from chance values; but c) baseline values and simulated values did not differ (see table 3). The one exception was the case of OR (Opposite Responses), where the difference between baseline subject and chance values was sizable but did not quite reach significance ($p = 0.059$).

4.3 Discussion

When parameters were set appropriately, the model produced sequences that are, on all measures considered, distinguishable from chance but indistinguishable from those produced by human subjects. Thus the model provides a promising account of human behaviour on the random generation task. This positive assessment should not be accepted at face value, however, as the model contains many parameters, particularly within the schema biases. Indeed the freedom to adjust those biases is critical in capturing the human biases indicated by the various measures of randomness.

Table 2: Means (and standard deviations) of the six dependent measures for the subject data of Cooper et al. (2012), chance, and simulation results.

Dependent Measure	Cooper et al. (2012)	Chance Value	Simulated Value
R	0.962 (0.567)	1.840 (0.838)	0.904 (0.456)
RNG	0.300 (0.068)	0.243 (0.019)	0.294 (0.041)
RR	0.014 (0.023)	0.093 (0.036)	0.016 (0.020)
AR	0.259 (0.143)	0.209 (0.042)	0.269 (0.097)
OR	0.131 (0.070)	0.107 (0.025)	0.148 (0.055)
TPI	0.733 (0.196)	0.956 (0.147)	0.711 (0.128)

Consider first R, equality of response usage. The chance value on this measure (1.840) reflects the fact that in a sequence of 100 independent responses chosen from a response set of 10 it is likely that not all responses are selected exactly 10 times. (If they were, R would be zero.) Human subjects consistently score below the chance value on this measure, indicating (perhaps counter-intuitively) that they produce responses with frequencies that are more equal than would be expected by chance in a 100 item sequence. This suggests not merely that human responses are not independent, but that response selection is somehow biased to actively avoid over-sampling of any particular response. This in turn requires either a strategy with a side-effect of equating response usage (e.g., cycling from one response to the next) or some memory of previous responses (or at least of response frequencies).

Consider now the various bigram biases. In order to achieve human-like values of RR, OR, AR and TPI within the model, the probability of *Task Control* selecting the "repeat" schema was set to 0.008, while for selecting an "opposite" schema it was 0.164, the probability of selecting the "adjacent clockwise" schema was set to 0.140 and for the "adjacent counter-clockwise" schema it was 0.113. (One might assume chance to be 0.100 for each schema.) While these values were set by hand to produce a good fit to the data, as noted above it is assumed that the values of the dependent measures arising from the human data reflect biases induced by the specific procedure or apparatus. Thus, the clock-face type display from which responses were selected appears to encourage high rates of selection of opposite and adjacent associates, just as two-handed keyboard response collection facilitates alternation in responses between hands. Departing from equi-probable schema selection within the model in order to address response bigram biases is therefore justified.

At the same time, if between-subject noise is not introduced into schema biases, the above-quoted settings for biases yield a value of RNG that is substantially lower than observed in human subject group data. To understand why consider two subjects, one with a tendency to favour production of adjacent responses and the other with a tendency to favour production of opposite responses. The first will yield a high score on AR and a low score on OR, while the second will yield the opposite pattern. Both subjects will however score similarly high on the RNG measure. The mean across subjects for AR or OR will therefore be moderate, while for RNG it will be high. It is therefore critical when considering group

Table 3: Pair-wise comparisons between the three datasets of table 2, with t-values (d.f. = 70 in all cases), and two-tailed probabilities.

Dependent Measure	Cooper et al. (2012) *versus* Chance Value	Simulated Value *versus* Chance Value	Cooper et al. (2012) *versus* Simulated Value
R	$t(70) = 5.203, p < 0.001$	$t(70) = 5.886, p < 0.001$	$t(70) = 0.482, p = 0.631$
RNG	$t(70) = 4.925, p < 0.001$	$t(70) = 6.828, p < 0.001$	$t(70) = 0.526, p = 0.601$
RR	$t(70) = 11.097, p < 0.001$	$t(70) = 11.261, p < 0.001$	$t(70) = 0.426, p = 0.671$
AR	$t(70) = 2.018, p = 0.047$	$t(70) = 3.421, p = 0.001$	$t(70) = 0.357, p = 0.722$
OR	$t(70) = 1.920, p = 0.059$	$t(70) = 4.074, p < 0.001$	$t(70) = 1.139, p = 0.258$
TPI	$t(70) = 5.440, p < 0.001$	$t(70) = 7.540, p < 0.001$	$t(70) = 0.582, p = 0.562$

random generation data to consider (and as in the case here, model) individual differences – some group measures of randomness may obscure critical features of individual behaviour (cf. Sexton & Cooper, 2014).

The repeat response (RR) measure is somewhat different from the other bigram measures. Both real and simulated subjects tend to avoid repeats, resulting in a significantly lower value for this measure than would be expected by chance. Towse (1998; see also Towse et al., 2016) suggests that this may be a result of automatically inhibiting a response once it has been generated, as in the Competitive Queueing mechanism of sequential control introduced by Houghton (1990). The model presented here instead assumes that low rates of repeat responses are the result of active monitoring and subsequent inhibition of responses that, according to an individual subject's conception of randomness, are insufficiently random. Thus repeat responses are proposed (albeit at low rates) by the *Apply Set* process but then suppressed by *Monitoring* prior to production by the *Generate Response* process. The rationale for this is discussed further below in the General Discussion.

The final measure, the turning point index (TPI) is again lower than chance in both real and simulated subjects. This reflects a tendency to favour one direction (clockwise) over the other. Again this reflects properties of the apparatus for response collection (i.e., the clock-face display), but again it is directly addressed within the model by assuming larger schema biases for clockwise selection than for equivalent counter-clockwise selection (e.g., as noted above, the adjacent clockwise schema has bias 0.140, while the adjacent counter-clockwise schema has bias 0.113).

Before turning to further empirical support for the model it is instructive to consider the extent to which the model can account for the five "robust behavioural effects" enumerated in the introduction. Most clearly, modality induced response biases (effect 2) may be attributed within the model to the set of available schemas, and their weights, which determine responses. Thus, if one were modelling random generation behaviour as reflected by responses from a two-handed key pad, one would assume that schemas for alternation would be available and strongly weighted. The effect of external representation of the response set (effect 1) may be due to similar factors, as the format of the external representation will likely impose variable schema-selection biases and thereby encourage stereotypy. This may also be the case for the effect of response set size (effect 4), where

larger response sets will require more schemas for generating a response from earlier behaviour if the full response set is to be produced (and so underlying biases are likely to be more pronounced). The relative lack of repeat responses (effect 3) arises in the model not due to inhibition of immediate responses (which must be overcome to produce a repeat), but due to a combination of a low rate of proposal of repeat responses together with deliberate suppression of repeats, with the latter being compromised when working memory is limited or monitoring for sufficient randomness is impaired. Finally, the effect of response rate (effect 5, with increased rate resulting in increased stereotypy) is a natural consequence of the combination of schema-generated behaviour and the monitoring / checking process. If the latter is compromised due to time pressure then responses will necessarily be more stereotyped.

5 Simulation 2: Experiment 1 of Cooper et al. (2012)

The target data for simulation 1 was from the control or baseline condition of the random generation study of Cooper et al. (2012). Simulation 2 explores whether the model can account for the random generation behaviour observed in the three different secondary task conditions of that study. We show that a) with appropriate values of key parameters the model can account for the behaviour of subjects in the different experimental conditions, but that b) the model is less able to capture artificial datasets, thus demonstrating that the model's parameters do limit or constrain its observable behaviours. Furthermore, we show that c) reduced forms of the model, with fewer parameters provide statistically less adequate accounts of the data. We interpret the effects of concurrent performance of the different secondary tasks on primary task performance in terms of the required parameter values in each condition. This allows us to infer how different secondary tasks compromise the various processes involved in random generation.

The logic of this simulation study assumes that the values of the various parameters generally reflect the efficacy of different cognitive processes. For example, it is assumed that concurrent performance of a secondary task will reduce the rate of switching between schemas (i.e., *Switch Rate* will be less than 1.00). Similarly it may reduce the efficiency of the monitoring process, such that it is applied on fewer trials than otherwise (i.e., *Monitoring Efficiency* may be less than 0.65), or reduce the temperature of the selection processes, such that schema selection is more biased (i.e., *Temperature* may be less than 1.00).

5.1 A 3-Parameter Model

Attempting to account for behaviour in terms of the full model, potentially varying all space of the key parameters, is both computationally impractical (requiring search in a five dimensional space) and theoretically dubious (given the large number of potentially freely varying parameters). In order to address both of these concerns we consider a reduced model in which two of the five key parameters are held fixed. Thus, we assume that it is not necessary to vary parameters specifically related to working memory (*WM Decay Time* and *WM Update Efficiency*) in order to simulate secondary task interference. As argued in relation to simulation 1, there is an inverse relation between *WM Decay Time* and *Monitoring Efficiency*, and so it is not necessary from a simulation perspective to consider variation of both (though it may be appropriate to consider both when interpreting simulation results). Moreover, as shown in the supplementary materials, the model's behaviour is very sensitive to *WM Update Efficiency*, with repeat responses dominating when *WM Update Efficiency* is decreased from 1.00. This does not mirror subject behaviour in any condition, and so this parameter is fixed at 1.00 for all simulations reported in this section.

Table 4: Best-fitting values of the three parameters and the resultant fits for the baseline and three dual-task conditions of Cooper et al. (2012).

Condition	Temperature	Switch Rate	Mon. Eff.	Fit (Max SD)
Baseline	0.90	1.00	0.58	0.09
Digit-Switching	0.45	0.71	0.00	0.19
2-Back	0.28	0.61	0.00	0.16
Go / NoGo	0.49	0.85	0.35	0.12

5.1.1 Method

Given that our aim was to determine parameter values yielding the best fit of the model to different data sets, a brute force approach was taken whereby the reduced model's output at each point in a 3 dimensional grid of parameter space was scored for the six dependent variables considered in the original empirical work. To ensure good estimates of the model's behaviour across the parameter space 360 sequences, each of 100 items, were generated at each point in the grid, corresponding to 360 virtual subjects, and mean values of each dependent measure were calculated and recorded for each point, yielding the DV "profile" for that point. The grid itself was defined by *Temperature* ranging from 0.00 to 2.00, *Switch Rate* ranging from 0.00 to 1.00, and *Monitoring Efficiency* ranging from 0.00 to 1.00, each in increments of 0.01. It thus comprised a total of 201 × 101 × 101 = 2,050,401 points.

The fitting procedure minimised the difference between observed and simulated behaviour on all six dependent measures simultaneously. Thus, for each of the baseline and three experimental conditions, the fit between model DV profile and subject DV profile was calculated at each point in parameter space as the maximum, over the six dependent measures, of the absolute value of the difference in *z* scores between the model's behaviour (based on 360 virtual subjects) and the subject behaviour on that condition. The values of the parameters for which this measure was minimised were taken as the best fitting values for that condition.

In order to determine whether the model was limited in the behaviours that it could produce, best fits were also calculated for two kinds of "artificial" profiles: a) the profile generated by a true random generation process (as given in the Expected Value column of table 1), and b) profiles generated by randomly crossing the actual profiles from the different experimental conditions. The former allows one to evaluate whether the model can fit a true random sequence while the latter allows one to evaluate whether the model can fit an arbitrary DV profile where all values in the profile are within the observed ranges.

5.1.2 Results

Best fitting parameter values and resultant fits for the baseline and three experimental conditions of Cooper et al. (2012) are shown in table 4. Thus, for the baseline condition the best fit (with all six dependent measures lying within 0.09 of a standard deviation of their observed values) was obtained with *Temperature* = 0.90, *Switch Rate* = 1.00, and *Monitoring Efficiency* = 0.58. These values compare favourably with those chosen by hand for simulation 1 (1.00, 1.00, 0.65, respectively), but also demonstrate that the good fit to baseline behaviour presented in that simulation is not the result of cherry-picking parameter values. In fact good fits to the baseline behaviour (within 0.20 standard deviations on all six dependent measures)

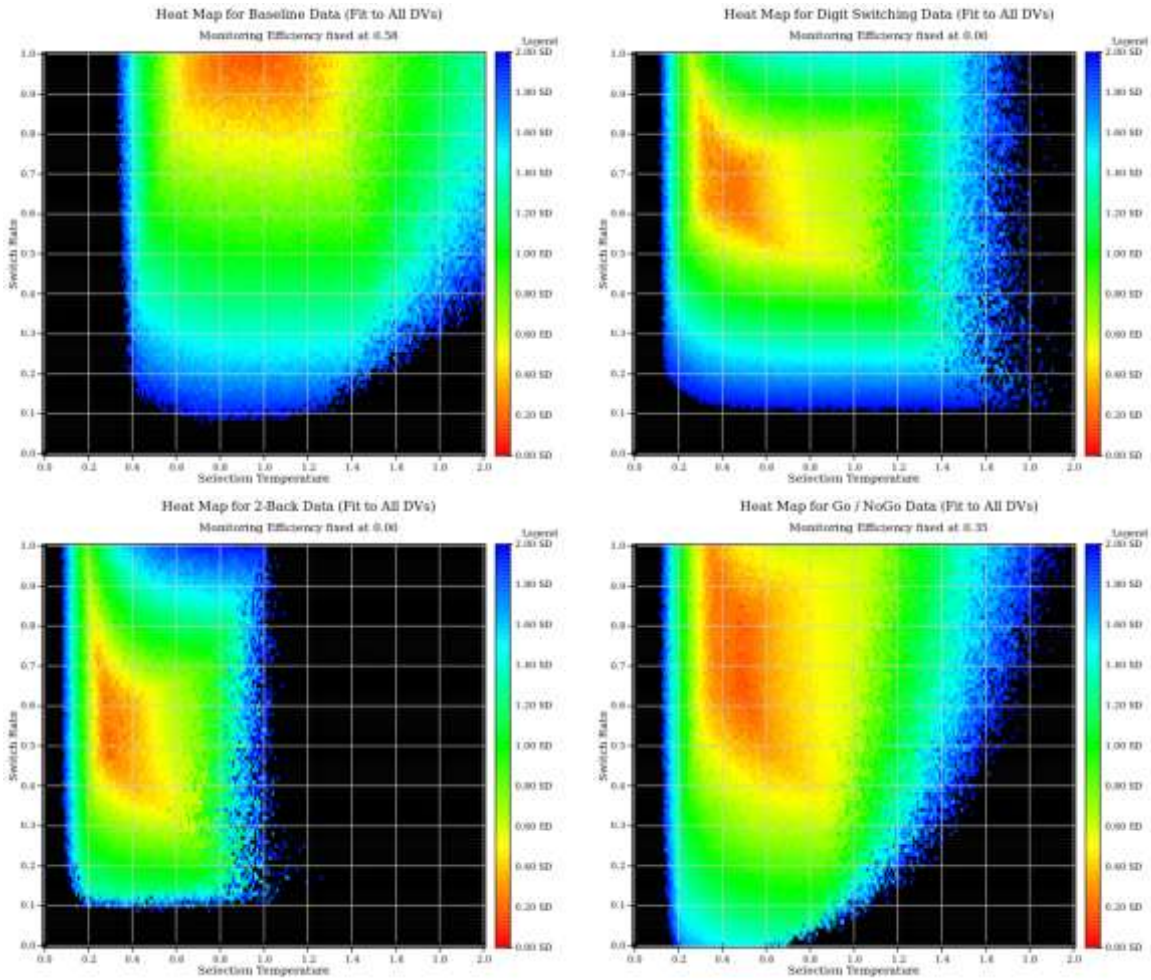

Figure 3: Heat maps showing fits for each condition of Cooper et al. (2012). Shades of red indicate fits within 0.2 standard deviations on all dependent measures, while black indicates a fit of more than 2.0 standard deviations on at least one dependent measure.

are obtained with any value of *Temperature* in the range [0.90, 1.00] and any value of *Monitoring Efficiency* in the range [0.55, 0.65].

Figure 3 shows heat-maps for cross-sections through the parameter space for each of the four conditions, with *Monitoring Efficiency* fixed in each case at the values given in table 4. In all four cases a good fit is obtained within a contiguous region of parameter space. The best fitting parameter values reported in table 4 should therefore be viewed as approximate estimates of the relevant parameters under the various conditions. Note in particular that while the best fit for the *Digit-Switching* condition occurs when *Switch Rate* is 0.71 (well above the best-fitting value of 0.61 for the 2-*Back* condition), a good fit is also obtained for this condition with lower values of *Switch Rate* (e.g., when *Switch Rate* is 0.61, the measure of fit increases only slightly to 0.21).

The graphs in figure 4 show the observed values of the six dependent variables for each condition along with their values for the best-fitting simulation for each condition. As can be seen from the figure, the best-fitting simulations capture all qualitative effects. Thus, R is highly elevated in the *Digit-Switching* and *2-Back* simulation in comparison to the *Baseline* simulation, but only slightly elevated in the *Go / NoGo* simulation in comparison to the *Baseline* simulation. In contrast, RNG is more highly elevated in the *2-Back* simulation (in comparison to the *Baseline* simulation) than in the *Digit-Switching* and *Go / NoGo* simulations, where it is moderately elevated.

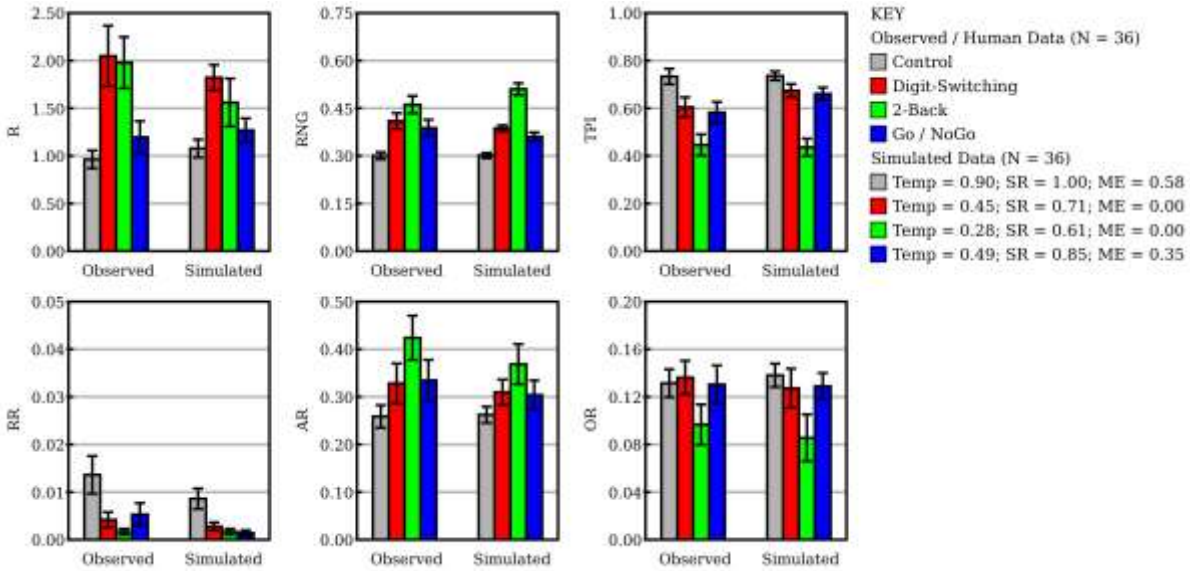

Figure 4: Simulation 2 results, comparing mean values for each dependent measure for human subjects in each experimental condition and best-fitting simulated subjects, with best-fitting parameter values determined as described in the text. Error bars represent one standard error.

With regard to the attempts to fit artificial profiles, the best fit to the true random profile was 2.03 (with *Temperature* = 1.97, *Switch Rate* = 0.99, and *Monitoring Efficiency* = 0.34). That is, the best fit of the model to a true random profile was more than two standard deviations away from the target data on at least one dependent measure. Note though that better fits occur when *Temperature* is higher, and the model's performance approaches true chance as *Temperature* increases to infinity when *Switch Rate* is 1.00 and *Monitoring Efficiency* is 0.00. Attempting to fit the model to the crossed profiles revealed that the majority of such profiles (68% of 5000) could not be fit by the model to within 0.20 of a standard deviation. Thus for artificial profiles the best fits were generally poorer (and often substantially so) than for each of the empirically observed profiles, indicating the model is better able to fit human-generated than artificial profiles.

5.2 Model Comparison: Restricted Models

When *WM Update Efficiency* is held at its maximum value (of 1.0) and *WM Decay Time* is fixed at 30, as in the above simulations, the model effectively has 3 parameters and is being evaluated for a fit across the four conditions. While the obtained fits (in terms of maximum z scores across the six dependent measures) are held to be good, it is possible that the goodness of fit is largely a result of the number of parameters. Models with fewer parameters might provide statistically better fits. This issue may be addressed by comparing the fit of the 3-parameter model with the fits of 8 reduced models (i.e., 8 models with fewer parameters) using a measure of fit that penalises parameters whose inclusion does not provide a sufficiency improvement in fit. The Bayesian Information Criterion is one such measure.

5.2.1 Method

Eight reduced models were constructed from the full 3-parameter model by restricting parameters as follows:

- Model 2: *Temperature* fixed at 1.0
- Model 3: *Switch Rate* fixed at 1.0
- Model 4: *Monitoring Efficiency* fixed at 0.6
- Model 5: *Switch Rate = Monitoring Efficiency*

Table 5: The nine models, the number of parameters (*k*) in each model, and their Bayesian Information Criterion (BIC) scores.

Model	*k*	*BIC*
1. Unrestricted	3	1.490
2. *Temperature* fixed @ 1.0	2	13.671
3. *Switch Rate* fixed @ 1.0	2	5.629
4. *Monitoring Efficiency* fixed @ 0.6	2	8.198
5. *Switch Rate = Monitoring Efficiency*	2	8.053
6. *Temperature* fixed @ 1.0 and *Switch Rate* fixed @ 1.0	1	13.799
7. *Temperature* fixed @ 1.0 and *Monitoring Efficiency* fixed @ 0.6	1	18.538
8. *Monitoring Efficiency* fixed @ 0.6 and *Switch Rate* fixed @ 1.0	1	7.613
9. *Temperature* fixed @ 1.0 and *Switch Rate = Monitoring Efficiency*	1	16.740

- Model 6: *Temperature* fixed at 1.0 and
- Model 7: *Temperature* fixed at 1.0 and *Monitoring Efficiency* fixed at 0.6
- Model 8: *Monitoring Efficiency* fixed at 0.6 and *Switch Rate* fixed at 1.0
- Model 9: *Temperature* fixed at 1.0 and *Switch Rate = Monitoring Efficiency*

Models 2, 3 and 4 explore whether each of the three remaining parameters is necessary. Each of these models eliminates one effective parameter by holding it fixed at a value that, from simulation 1, is known to yield a near maximal fit for the baseline condition. Model 5, in contrast, explores the possibility that a single efficiency parameter, specifying both *Switch Rate* and *Monitoring Efficiency*, might be sufficient to account for the data. These four models (models 2, 3, 4 and 5) thus have 2 effective parameters. Models 6 to 9 combine the constraints in models 2 to 5, and have just one effective parameter.

For each of the nine models, the Bayesian Information Criterion was calculated using equation 4:

$$BIC = n + n \cdot \ln 2\pi + n \cdot \ln \frac{SSE}{n} + (k+1) \cdot \ln n \tag{4}$$

where n is the number of conditions, SSE is the minimum sum squared error between the model and the data across the four conditions, and k is the number of parameters in the model. SSE was calculated as the sum of squares of fits over the four conditions, where the fit for each condition was calculated as the minimum of the maximum z-score difference across the six dependent measures between the observed data and the simulated data (i.e., using figures as in table 4).

5.2.2 Results

Table 5 shows the number of parameters and BIC scores for each of the nine models. Model 1 is preferred on the grounds that, despite having more parameters than the other models, its BIC score is substantially lower than that of any other model. In other words, the extra degrees of freedom of model 1 given by the additional parameters in comparison to the other models is more than compensated for by the improvement in fit of model 1 in comparison to

the other models. It is also of note that the models in which *Temperature* is fixed (models 2, 6, 7 and 9) perform very poorly (all with BIC scores above 13). Moreover models in which *Monitoring Efficiency* is fixed perform more poorly than equivalent models in which *Switch Rate* fixed (i.e., model 4 versus model 3, model 7 versus 6). This suggests that a) variation in *Temperature* is more critical than variation in any other parameter for capturing the behavioural data, while b) variation in *Switch Rate* is less critical than variation in any other parameter for capturing the behavioural data.

5.3 Discussion

When allowed to vary the three key parameters of *Temperature*, *Switch Rate*, and *Monitoring Efficiency* arbitrarily, the model can provide a good fit to human random generation behaviour in all four conditions of Cooper et al.'s (2012) study. This may seem unsurprising given the number of parameters that have been varied, but given that fitting each condition requires simultaneously fitting the values of six dependent measures it is not a straightforward result. The facts that the model is also generally less able to fit various "artificial" profiles, and that reduced forms of the model with fewer parameters provide statistically poorer fits to the behavioural data, add further weight to this claim.

Modelling of secondary task interference suggests that the effect of a secondary task is not just to magnify inherent schema selection biases (modelled by a reduction in *Temperature*), but also to reduce both the rate of schema switching and the efficiency of monitoring processes (or equivalently, the persistence of items in working memory). The three secondary tasks of Cooper et al. (2012) differ in the impact that their concurrent performance has on each of these mechanisms. Thus, the *Go / NoGo* task is least disruptive, being well-modelled by a relatively moderate reduction in *Temperature* (from 1.00 to 0.49), a relatively small reduction in *Switch Rate* (from 1.00 to 0.85), and a sizable reduction in *Monitoring Efficiency* (from 0.65 to 0.35). Further reductions in each of the three parameters allows the effect of concurrent performance of the *Digit-Switching* task on random generation to be captured, though the additional reduction in *Temperature* is small (down to 0.45) and arguably unreliable, in comparison to the reduction in the other two parameters (down to 0.71 and 0.00, respectively). That the best fit is obtained with *Monitoring Efficiency* at 0.00 is considered below, though this is also found for the *2-Back* condition, which in addition requires a further substantial reduction in *Temperature* (down to 0.28) together with a further small reduction in *Switch Rate* (down to 0.61).

How should these results be interpreted? One possibility is that the three tasks vary on a single dimension, such as difficulty, with the *2-Back* task being hardest and hence having greatest effect on all three parameters (and with *Monitoring Efficiency* subject to a floor effect) and the *Go / NoGo* task being easiest and hence having least effect on all three parameters. The model supports a more nuanced account, however. First, all three secondary tasks appear to compromise monitoring of the primary task, though less so for the *Go / NoGo* task.[5] This is evidenced both by the best fitting parameters values across conditions from the full 3-parameter model and by the relatively poor fits of reduced models in which *Monitoring Efficiency* was fixed. Together, this evidence supports the idea of a separable monitoring process that, under dual-task conditions, must be shared by multiple tasks. Second, set-shifting or switching is compromised in all dual task conditions. Again, this is evidenced both by the best fitting parameter values in the 3-parameter model and by the relatively poor fits of

[5] Note that in the *Go / NoGo* task the no-go signal occurs on only one in six trials, in contrast to the other tasks, where monitoring is required on every trial. The fact that subjects, when completing the *Digit-Switching* task, frequently attempted to correct errors suggests that the task does involve monitoring at some level, and if one were to develop a process model of the *Digit-Switching* task, then in all likelihood it would need to include a monitoring function to capture this aspect of behaviour.

reduced models in which *Switching* was fixed. That switching might be compromised in dual task conditions is plausible if only because when dual-tasking one must frequently switch between the primary and secondary tasks. Third, the choice of which schema to switch to is also affected in all dual task conditions – models in which the Temperature parameter is fixed at a value that yields a good fit in the baseline condition perform poorly when evaluated across the full dataset – though this choice appears to be particularly compromised when simultaneously performing the 2-Back task.

Note that this account does not make explicit appeal to a 'response inhibition' executive function. The original empirical work assumed that the *Go / NoGo* task would (primarily) tap this function, based in part on the theoretical perspective of Miyake et al. (2000). There is mounting evidence to question whether response inhibition is in fact a distinct executive function (e.g., Hampshire & Sharp, 2015; Miyake & Friedman, 2012), and the current model does not include a specific response inhibition function (though inhibition of putative responses does occur when the *Monitoring* process detects a putative response that is insufficiently random; cf. figure 2).

Monitoring Efficiency concerns rejecting individual responses in the context of (memory of) previous responses. The model accounts for the effect of concurrent performance of a secondary task on random generation by assuming that recent responses are reliably encoded in a short-term store but that (one form of) variation from baseline performance arises through intermittent failure of monitoring of putative responses with reference to this store. When monitoring is completely disabled, random generation reduces to intermittent switching between schema-based responses, with the difference in interference patterns resulting from differences in the rate of switching and the strength of bias in choosing schemas when switching.

With regard to switching, the *2-Back* task appears to affect *Switch Rate* more than the *Digit-Switching* task, with it being reduced to 0.61 for the best fitting model in the former condition and 0.71 in the latter condition. While this may seem counter-intuitive, it is unclear whether the difference is statistically significant – reasonable fits in both conditions can be obtained with an intermediate value.[6] However, it is noteworthy that in the empirical study *Digit-Switching* required that subjects switch stimulus-response mappings on every four trials, and not on every trial. In other words, the *Digit-Switching* task was not maximal in its set-shifting requirements. Moreover, while the *2-Back* task is normally viewed as a task that draws on working memory monitoring and maintenance, if working memory maintenance is achieved through a rehearsal process, then successful performance in the *2-Back* condition may involve frequent shifting between this rehearsal process and the primary task, thus compromising switching within the primary task. To explore the viability of this hypothesis it will be necessary to integrate a complete process model of the *2-Back*, and other, secondary tasks with the model of random generation. One consequence of this would be clarification of the extent to which the three secondary tasks are "process pure" (i.e., the extent to which each taps one and only one executive function). Additional empirical work, in which the switching frequency of the *Digit-Switching* task were varied, would also speak to this hypothesis.

Interpretation of the *Temperature* parameter, and how it is affected by the different secondary tasks, is more complex, partly because its affects are determined by an equation rather than a complete process account of schema selection biases. The parameter is clearly important, however, as shown by its importance in the model comparison study. We consider the interpretation of *Temperature* in more detail in the General Discussion.

[6] Fixing *Switch Rate* at 0.66 and *Monitoring Efficiency* at 0.00 results in a best of 0.22 for the *Digit-Switching* condition (up from 0.19) and 0.24 for the *2-Back* condition (up from 0.16). In both cases the fit remains low.

6 General Discussion

We have presented a computational cognitive model of subject performance on the random generation task. The model operationalises and extends the verbal model originally proposed by Baddeley et al. (1998), placing it within the framework of the Contention Scheduling / Supervisory System architecture of Norman and Shallice (1986). It provides a good account not only of performance on the basic task (simulation 1), but also of the operation and interaction of executive or control processes involved in the task (simulation 2). More specifically, modelling of concurrent task performance suggests that all three secondary tasks compromise the functions of set-shifting, monitoring and schema selection, with the different interference patterns arising from different secondary tasks being accounted for by different degrees of compromise to these functions. In this section we consider two issues in depth: the relationship between model parameters and executive functions (and the origins of various well-known biases in terms of these parameters and functions), and the model as an instance of a task within the general Contention Scheduling / Supervisory System architecture.

6.1 Parameters and their relation to Executive Functions

6.1.1 The control of Working Memory

A critical feature of the model is the hypothesised involvement of specific parameterised functions which are motivated by the executive functions literature. Consider first the operation of working memory. This is implemented as a buffer with a decay function. The "updating and monitoring of working memory representations" is one of the three executive functions considered in detail by Miyake et al. (2000; see also Miyake & Friedman, 2012). The structural equation modelling presented there suggests that this function is critical to maintaining equality of response usage. The current model supports this suggestion in that the R statistic increases as the efficiency of the monitoring process decreases (cf. supplementary materials, figure S1, central panel).

The model further demonstrates, however, that R is also sensitive both to the updating of information in working memory and the *maintenance* of that information. Thus R increases both as *WM Update Efficiency* decreases and as *WM Decay Time* decreases (cf. figure S1, upper two panels), though the former case also leads to higher than observed values of repeat responses (RR). A large proportion of the executive functions literature related to the operation of working memory concentrates on updating (e.g., Collette et al., 2005; Reineberg et al., 2015; Sharinger et al., 2015) or maintenance (e.g., Ashby et al. 2005; Unsworth & Engle, 2007), or both (e.g., O'Reilly & Frank, 2006), rather than monitoring (but see: Shallice et al., 2008; Shallice & Cooper, 2011). A consequence of compromising updating within the model is the production of a high number of repeat responses – more than is observed in the data of Cooper et al. (2012), where repeat responses were found to be even less frequent under dual-task conditions than under single-task conditions. The model here thus argues in favour of maintenance or monitoring and against updating, as the critical working memory control function(s) in random generation.

6.1.2 The Temperature parameter, cognitive load and meta-monitoring

The second executive function of Miyake et al. (2000) held to influence random generation – inhibition of prepotent responses – is not explicitly included in the model. Effects held by Miyake and colleagues to be related to this function are instead accounted for through the process of schema selection, and in particular through the *Temperature* parameter. When *Temperature* is low, biases in the prepotency of individual schemas are strong, so the model tends to select those schemas that are most prepotent (e.g., opposite or adjacent). As *Temperature* increases, all schemas become equi-probable, meaning that schemas which might otherwise have low-probability of selection are just as likely to be selected as prepotent

schemas. More generally the *Temperature* parameter may be conceived of as regulating strategy generation and reflecting cognitive load. At higher temperatures (e.g., when not performing a secondary task and when cognitive load is moderate), the model is more able to select schemas that fulfil task requirements (i.e., that reflect an appropriate strategy for the task), though biases are still present (e.g., the "repeat" schema is chosen relatively rarely). At lower temperatures (e.g., when performing a demanding secondary task, such as the 2-back task), the model's ability to select task-appropriate schemas is compromised and the model falls back on selecting prepotent schemas. Schemas that are rarely selected at the best of times, such as "repeat", are consequently even less likely to be selected.

An alternative way of conceiving of *Temperature* is in terms of meta-monitoring. The random generation task requires that one avoid selection of prepotent schemas. At the algorithmic level, this might be achieved through a cyclic process that proposes a schema, evaluates it against some criterion (e.g., novelty), and rejects it if it fails that criterion, leading to proposal of an alternative schema, and so on until either the proposed schema passes the criterion or a time limit forces acceptance of an otherwise disfavoured schema. This conception can also be related to the less well-defined notion of cognitive load (high load will reduce the ability to cycle through the process of schema selection), as well as the effects of rate on random generation discussed earlier (faster production limits the schema selection process, leading to selection of prepotent schemas and hence the production of more stereotyped responses, or alternatively when forced to make a fast choice one selects the first thing that comes to mind), and the effects of TMS over left DLPFC reported by Jahanshahi et al. (1998, i.e., that TMS over this region results in increased selection of stereotyped counting up / down schemas and consequent decreased selection of other schemas such as "counting by 2").

These alternative conceptions of *Temperature* are consistent with the account given here because schema selection is specified at an abstract level in terms of an equation rather than an algorithm. This abstract level is appropriate for modelling random generation as the internal details of the selection process are not required in order to account for the behavioural data. Any algorithm for schema selection that is consistent with equation 3 will be able to account for the target behaviours considered in this paper.[7] Indeed, the behavioural data arguably cannot distinguish between these algorithmic specifications of the selection process and hence specifying the process at an algorithmic level would be unjustified.

6.1.3 Switching

The final parameter of the model, *Switch Rate*, determines how frequently schemas are deselected. When switch rate is low, the model tends to persist with a single schema over multiple responses. This leads to a reduction in the Turning Point Index (TPI) of the sequences generated by the model. While it might be argued that this parameter reflects task-specific strategic factors rather than executive control, the fact that TPI is systematically influenced by concurrent performance of a secondary task suggests that *Switch Rate* does fall within the remit of executive function. The link is presumably offered by the set-shifting function, on the assumption that set-shifting is equivalent to deselecting one schema and selecting another. Set-shifting is not considered to be a critical executive function in random generation by Miyake et al. (2000). Nor is it emphasised in the verbal model of Baddeley et al. (1998), which focuses on the similarities of the random generation task to the verbal fluency task rather than the differences. In verbal fluency subjects use a schema to generate

[7] Another possible algorithm would be an activation-based one with biases on schemas such that prepotent ones are more likely to reach threshold first, but with noise so that even those schemas with low bias could be selected, and with noise related to temperature such that high temperature corresponds to high noise (and hence more random selection) and low temperature corresponds to low noise (and hence more deterministic selection).

multiple associated responses (e.g., farmyard animals, when enumerating animals). In verbal fluency set-shifting is only required when the associate responses are exhausted. In contrast random generation requires shifting as frequently as possible. Two related hypotheses are forthcoming from the account presented here: first, the executive function of set-shifting is also critical for random generation, though its effect is primarily on dependent variables that specifically index schema switching (such as TPI), and hence second, secondary tasks that compromise set-shifting (such as the *Digit-Switching* task) will affect indices of schema switching.

6.1.4 Individual differences in executive functioning

A comment is also in order on modelling individual differences. This has been achieved in the current work by assuming individual differences in schema selection biases (cf. Appendix B). This is not intended to deny the effect of individual differences in executive functioning on random generation. Such individual differences might be captured by assuming that there is a small amount of individual variability in the parameters considered here.

6.2 The Contention Scheduling / Supervisory System architecture

In developing the model a key consideration was the relationship between the various putative processes involved in random generation and the wider cognitive architecture assumed to support general intelligent behaviour. Developing a model within a cognitive architecture has the advantage of placing general constraints on the model – constraints that flow from the need for the architecture to support behaviour across a range of tasks (Newell, 1990).

Many cognitive architectures have been proposed over the last 25 years, and several of them are considerably more developed (in computational terms) than the specific cognitive architecture – the Contention Scheduling / Supervisory System (CS/SS) architecture – adopted here. It is unclear, however, either how the various processes proposed in the original Baddeley et al. (1998) verbal model might fit within such architectures, or how such architectures might relate to the executive processes assumed to be involved in regulating behaviour on random generation tasks. Consider the ACT-R architecture (Anderson, 1993, 2007), which has been successfully applied across a wide range of cognitive tasks. ACT-R consists of a central serial production system that coordinates the inputs and outputs of a range of peripheral modules which take care of long-term memory and modality-specific processing. It does not consider routine schema-driven behaviour as distinct from deliberative behaviour. Nor does it consider abstract processes such as monitoring, how such processes might effectively run in parallel with the serial production system, or how parallel processes might be regulated by other executive functions (though see Altmann & Gray, 2008, for an account of task switching within an ACT-R framework). The Soar cognitive architecture (Newell, 1990; Laird, 2012) is possibly more promising in this sense. Soar, which has its roots in work on human problem solving, can be argued to draw a distinction between routine and non-routine (or in the language of Soar, “impasse-driven”) behaviour. Routine behaviour is driven by production rules that propose a preferred response in the current context. When no single response is preferred (which in the context of random generation may be because the otherwise preferred response is insufficiently random), an impasse arises and the architecture generates a subgoal to resolve that impasse. Arguably what Soar lacks in comparison to the CS/SS architecture is the idea that monitoring is a domain-general concurrent or parallel process that operates across cognition.

A significant advantage that architectures such as Soar and ACT-R (currently) have over the CS/SS architecture is that they are implemented as complex computer environments that take task specifications and enact them. At present there is no domain-general reference implementation of the CS/SS architecture, and while models of a handful of tasks have been

developed within the broad CS/SS theoretical perspective (e.g., routine sequential "multiple-object" tasks: Cooper, 2007; and the Wisconsin Card Sort Test: Sood & Cooper, 2013), further work in this area is needed. We view the model of random generation presented here as a step in this direction. Moreover while the diagrammatic rendition of the model within the CS/SS architecture (figure 1) is based on the broad functional decomposition of supervisory processes described by Shallice and Burgess (1996), we envisage substantial decomposition of the existing components, as well as specification of additional functional components. Thus, *Task Control* is highly underspecified, while the motor control literature (e.g., Wolpert & Ghahramani, 2000) suggests that the architecture could reasonably be augmented with a forward model of anticipated consequences of actions, which is then used (by *Monitoring*) to compare anticipated sensory feedback with actually sensory feedback in order to detect errors. Nevertheless, the current decomposition of processes encapsulated in the CS/SS architecture provides insight into how individual tasks might be specified within the architecture. That insight includes distinguishing between monitoring and task control, and the use by task control of basic schemas which must be activated or selected (by task control) in sequence in order to achieve task goals.

6.3 Conclusion

We have presented a process model of the random generation task that is embedded within a cognitive architecture that is itself motivated by work on executive function. As such, the work presented here demonstrates at the algorithmic level how random generation draws upon multiple executive functions. It includes several key insights concerning the random generation task. First, it clarifies how equality of response usage may be maintained via a record of recent responses (in working memory). Empirical data indicate that when that record is reasonably accurate, human-generated sequences can, paradoxically perhaps, achieve more equal response usage than would be expected by chance. More chance-like behaviour results when that record is degraded or ineffectively monitored. Second, it is shown that model can account for human performance on a random generation task as scored on multiple indices of random generation, and how those indices are affected by concurrent performance of a secondary task (and arguably also by variables such as generation rate and by lesions or TMS to left DLPFC). Schema selection biases play an important role in this account. Third, it suggests a critical role for set-shifting in random generation – an aspect of random generation that has largely been ignored in previous considerations of the task.

The account also includes preliminary insights concerning the integration of executive functions within a coherent cognitive architecture. On the one hand, previous computational models of executive function have largely focused on one or two executive functions, without considering those functions in the wider context of a complete cognitive architecture. On the other hand, previous cognitive architectures have largely been developed based on consideration of general problem solving or the performance of complex tasks. In combining these two domains of enquiry, we have fleshed out the CS/SS theory to a computationally complete cognitive architecture in which key executive functions have distinct and explicit roles in regulating task behaviour.

Acknowledgements

I am grateful to Matt Jones, Erik Altmann and one further (anonymous) reviewer for constructive comments on an earlier draft of this manuscript, and to Nick Sexton for discussions concerning the model.

References

Altmann, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological Review*, *115*, 602-639.

Anderson, J. R. (1993). *Rules of the Mind*. Lawrence Erlbaum Associates.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Aron, A.R., Robbins, T.W., & Poldrack, R.A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*, 177-185.

Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). FROST: A distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*, 1728-1743.

Baddeley, A. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology*, *18*, 119-129.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *49*, 5-28.

Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998). Random generation and the executive control of working memory. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *51*, 819-852.

Band, G. P., Van Der Molen, M. W., & Logan, G. D. (2003). Horse-race model simulations of the stop-signal procedure. *Acta Psychologica*, *112*, 105-142.

Boucher, L., Palmeri, T. J., Logan, G. ., & Schall, J. D. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review*, *114*, 376-397.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, *55*, 37-85.

Brugger, P. (1997). Variables that influence the generation of random sequences: An update. *Perceptual and Motor Skills*, *84*, 627-661.

Collette, F., Van der Linden, M., Laureys, S., Delfiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Exploring the unity and diversity of the neural substrates of executive functioning. *Human Brain Mapping*, *25*, 409-423.

Cooper, R. P. (2007). Tool use and related errors in ideational apraxia: The quantitative simulation of patient error profiles. *Cortex*, *43*, 319-337.

Cooper, R. P. (2010). Cognitive control: Componential or emergent? *Topics in Cognitive Science*, *2*, 598-613.

Cooper, R. P. & Fox, J. (1998): COGENT: A visual design environment for cognitive modelling. *Behaviour Research Methods, Instruments, & Computers*, *30*, 553-564.

Cooper, R. P., & Shallice, T. (2000). Contention Scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297--338.

Cooper, R. P., Wutke, K., & Davelaar, E. J. (2012). Differential contributions of set-shifting and monitoring to dual-task interference. *Quarterly Journal of Experimental Psychology*, *65*, 587-612.

Dennett, D. C. (1998). Reflections on language and mind. In P. Carruthers & J. Boucher (eds.), *Language and Thought: Interdisciplinary Themes* (pp. 284-295). Cambridge University Press, Cambridge, UK.

Evans, F. J. (1978). Monitoring attention deployment by random number generation: An index to measure subjective randomness. *Bulletin of the Psychonomic Society*, *12*, 35-38.

Gilbert, S. & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*, 297-337.

Hampshire, A. & Sharp, D. J. (2015). Contrasting network and modular perspectives on inhibitory control. *Trends in Cognitive Sciences*, *19*, 445-452.

Herd, S. A., Hazy, T. E., Chatham, C. H., Brant, A. M., & Friedman, N. P. (2014). A neural network model of individual differences in task switching abilities. *Neuropsychologia*, *62*, 375-389.

Jahanshahi, M., Dirnberger, G., Fuller, R. & Frith, C. D. (2000). The role of dorsolateral prefrontal cortex in random number generation: A study with positron emission tomography. *Neuroimage*, *12*, 713-725.

Jahanshahi, M., Profice, P., Brown, R. G., Ridding, M. C., Dirnberger, G., & Rothwell, J. C. (1998). The effects of transcranial magnetic stimulation over the dorsolateral prefrontal cortex on suppression of habitual counting during random number generation. *Brain*, *121*, 1533-1544.

Jahanshahi, M., Saleem, T., Ho, A. K., Dirnberger, G., & Fuller, R. (2006). Random number generation as an index of controlled processing. *Neuropsychology*, *20*, 391-399.

Laird, J. E. (2012). *The Soar Cognitive Architecture*. The MIT Press, Cambridge, MA.

Miyake, A. & Friedman, N. P. (2012). The nature and organisation of individual differences in executive function: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8-14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49-100.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*, 134-140.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. *Cerebral Cortex*, *23*, 264-282.

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, and D. Shapiro (eds.) *Consciousness and Self Regulation, Volume 4* (pp. 1-18). Plenum: NY.

O'Reilly, R. C. & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*, 283-328.

Peters, M., Gieshrecht, T., Jelicic, M, & Merckelbach, H. (2007). The Random Number Generation Task: Psychometric properties and normative data of an executive function task in a mixed sample. *Journal of the International Neuropsychological Society*, *13*, 626-634.

Proios, H., Asaridou, S. S. & Brugger, P. (2008). Random number generation in patients with aphasia: A test of executive functions. *Acta Neuropsychologica*, *6*, 157-168.

Reineberg, A. E., Andrews-Hanna, J. R., Depue, B. E., Friedman, N. P., & Banich, M. T. (2015). Resting-state networks predict individual differences in common and specific aspects of executive function. *NeuroImage*, *104*, 69-78.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*, 358-367.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology*, *52*, 1293-1304.

Sexton, N. J., & Cooper, R. P. (2014). An architecturally constrained model of random number generation and its application to modelling the effect of generation rate. *Frontiers in Psychology*, *5*, 1-14.

Shallice, T. & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*, 1405-1412.

Shallice, T., & Cooper, R. (2011). *The Organisation of Mind*. 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.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Sood, M., & Cooper, R. (2013). Modelling the supervisory system and frontal dysfunction: an architecturally grounded model of the Wisconsin Card Sorting Task. In M. Knauff, N. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1354-1359). Austin, TX: Cognitive Science Society.

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*, 77-101.

Towse, J. N. & Neil, D. (1998). Analyzing human random generation behaviour: A review of methods used and a computer program for describing performance. *Behaviour Research, Instruments and Computers*, *30*, 583–591.

Towse, J. N., Towse, A. S., Saito, S., Maehara, Y. & Miyake, A. (2016). Joint cognition: Thought contagion and the consequences of cooperation when sharing the task of random sequence generation. *PLOS One*, *11*(3), e0151306. doi:10.1371/journal.pone.0151306

Towse, J. N. & Valentine, J. D. (1997). Random generation of numbers: A search for underlying processes. *European Journal of Cognitive Psychology*, *9*, 381-400.

Tunes, G. S. (1964). A brief survey of variables that influence random generation. *Perceptual and Motor Skills*, *18*, 705-710.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104-132.

Vandierendonck, A., Liefooghe, B. & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*, 601-626.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*, 65-72.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*, 329-355.

Wiegersma, S. (1984). Forward and backward continuations in produced number sequences. *Perceptual and Motor Skills*, *58*, 735-741.

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*, 1212-1217.

Appendix A: The Complete Rule-Set of the Model

Apply Set

If there is no recollection of a previous response (e.g., on the first trial), act at random

IF *Response Buffer* is empty and
Working Memory is empty
THEN add response(R, T_c) to *Response Buffer*
WHERE R is a random number in [0, 10) and
T_c is the current time

Generate a putative response based on the most recent response and the current schema

IF *Response Buffer* is empty
response(R_1, _) is in *Working Memory* and
schema(S) is in *Schema Network*
THEN add response(R_2, T_c) to *Response Buffer*
WHERE R_2 is the result of applying S to R_1 and
T_c is the current time

Generate Response

Output the putative response and retain a record of it in Working Memory

IF response(R, T_{c-2}) is in *Response Buffer*
THEN delete response(R, T_{c-2}) from *Response Buffer* and
say R and
add response(R, T_c) to *Working Memory*
WHERE T_{c-2} was the time two cycles ago and
T_c is the current time

Task Control

If no schema is selected, then select one at random

IF response(R, _) is in *Working Memory* and
schema(_, selected) is not in *Schema Network*
THEN delete schema(S, unselected) from *Schema Network*
add schema(S, selected) to *Schema Network*
WHERE S is randomly selected from the set of weighted schemas

Deselect (i.e., switch) the selected schema, ideally after every response

IF response(_, T_{c-1}) is in *Working Memory* and
schema(S, selected) is in *Schema Network*
THEN delete schema(S, selected) from *Schema Network*
add schema(S, unselected) to *Schema Network*
WHERE T_{c-1} was the time one cycle ago

Monitoring

If a putative response is insufficiently random, don't generate it and switch schemas

IF response(R, _) is in *Response Buffer* and
R is not subjectively sufficiently random
THEN delete schema(S, selected) from *Schema Network*
add schema(S, unselected) to *Schema Network*
delete response(R, _) from *Response Buffer*

Appendix B: Schemas and schema biases

As described in the main text, schema selection is determined by equation 3, which is a function of the bias or weight of each schema and τ, the *Temperature* parameter. For the clock-face type random generation task modelled here, we assume 11 schemas as shown in Table A1, though these should be understood as being proxies for the schemas actually used by subjects. That is we do not suppose that all subjects possess a "+3" schema or a "–2" schema. Moreover the specific schemas and biases will clearly be a function of the specific form of the random generation task. Thus if responses were produced on a two-handed keyboard, one would require a schema for generating the homologous response on the other hand, and this would presumably be strongly biased. The theoretically critical feature of the schema set is that it allows production of any response given any previous response.

Schema biases were set by hand to produce biases similar to those in the baseline condition of Cooper et al. (2012). Table A1 shows the values, together with the probability of selection given by equation 3 at three different values of *Temperature*. At low values, underlying biases are very strong (and the schema with the strongest bias dominates as *Temperature* approaches zero). At high values all schemas become equally likely. Note that it is also assumed that there are individual differences in biases. This is modelled by perturbing biases on a virtual-subject-by-virtual-subject basis. Individual values of biases were obtained by multiplying the biases in table A1 by a normally distributed random variable ($\mu = 1.0$, $\sigma = 0.5$), and clipping the result to be non-negative.

Table A1: Schema biases and the effect of the *Temperature* parameter. The table lists all schemas included in the model, their hard-coded bias, and the resultant probability of schema selection at three different values of *Temperature*.

Schema	Bias	p(selection) at *Temperature* = 0.5	p(selection) at *Temperature* = 1.0	p(selection) at *Temperature* = 2.0
Repeat	0.10	0.001	0.008	0.027
+1	1.80	0.193	0.140	0.116
+2	1.40	0.117	0.109	0.102
+3	1.35	0.108	0.105	0.100
+4	1.40	0.117	0.109	0.102
+5	1.05	0.066	0.082	0.089
–5	1.05	0.066	0.082	0.089
–4	1.10	0.072	0.086	0.091
–3	1.05	0.066	0.082	0.089
–2	1.10	0.072	0.086	0.091
–1	1.45	0.125	0.113	0.104

Supplementary Materials

7 Effects of Varying Key Parameters

One issue that arises from simulation 1 concerns the degree to which the randomness or otherwise of the sequences that the model generates are dependent upon the values of its parameters. A more detailed examination of the model's parameter space is in order both to understand the mechanisms within the model that lead to deviations from expected values for the various measures and to ensure that the model cannot capture *any* pattern of behaviour (cf. Roberts & Pashler, 2000). This is the purpose of the simulation presented here. We consider the five key parameters relating to the operation of *Working Memory*, *Monitoring*, and *Task Control* discussed above, namely *WM Decay Time*, *WM Update Efficiency*, *Monitoring Efficiency*, *Switch Rate*, and *Temperature*.

7.1 Method

Five separate simulations were run, one for each of the key parameters. In each case, the value of the parameter was varied across a wide range, with 21 values chosen at equal intervals from the minimum to the maximum values. (For example, *WM Decay Time* was varied from 0 to 40 at intervals of 2, while *Switch Rate* was varied from 1.00 to 0.00 at intervals of 0.05.) Each simulation was run 36 times at each setting of the parameters to yield a sample of 36 simulated subjects for each point in parameter space. The means and standard deviations of all six indices of randomness were then calculated for each sample within each of the five simulations.

7.2 Results

Figure S1 shows how the values of the various measures of randomness (transformed to *z*-scores based on the mean and standard deviation of the baseline subjects) vary as a function of each parameter, when all other parameters are held at their values used in simulation 1. Panel a) of the figure shows the case of varying *WM Decay Time*. Several measures, particularly RNG, RR and AR, are insensitive to variation in *WM Decay Time*, while others – most notably R and TPI – are not. Thus when *WM Decay Time* is low (e.g., up to about 8 cycles), the model produces an R score that is one to two standard deviations above that of human subjects. As *WM Decay Time* increases, R decreases to the level of human subjects. TPI behaves in a similar fashion. The former reflects the role that a record of recent responses (held in *Working Memory*) plays in balancing response usage. *Monitoring* typically rejects putative responses that are known to have been produced recently (on the basis of their presence in *Working Memory*). Degrading *Working Memory* therefore results in an increasing R score, resulting ultimately in a score that is similar to chance when *WM Decay Time* is very low (4 or less). The high values of TPI when *WM Decay Time* is low also reflect chance performance on this measure. As one would expect, however, the deviation from baseline performance for all dependent measures is close to zero when *WM Decay Time* is near the value adopted in simulation 1 (i.e., 30 cycles).

Panel b) of the figure S1 shows the effect of decreasing *WM Update Efficiency* from its default value of 1.00 to its minimum value of 0.00. Again, the various measures of randomness are differentially sensitive to the manipulation, but RR and R are clearly highly sensitive to any departure from perfect updating. Indeed, the model only matches human performance on these measures when *WM Update Efficiency* is 1.00. The reason for the high rates of repeat responses at lower values is that when *WM Update Efficiency* is less than 1.00 there is a possibility that responses are generated without being recorded in *Working Memory*. In the absence of such a record, *Monitoring* has no way to inhibit putative responses (like repeats) that would normally be suppressed.

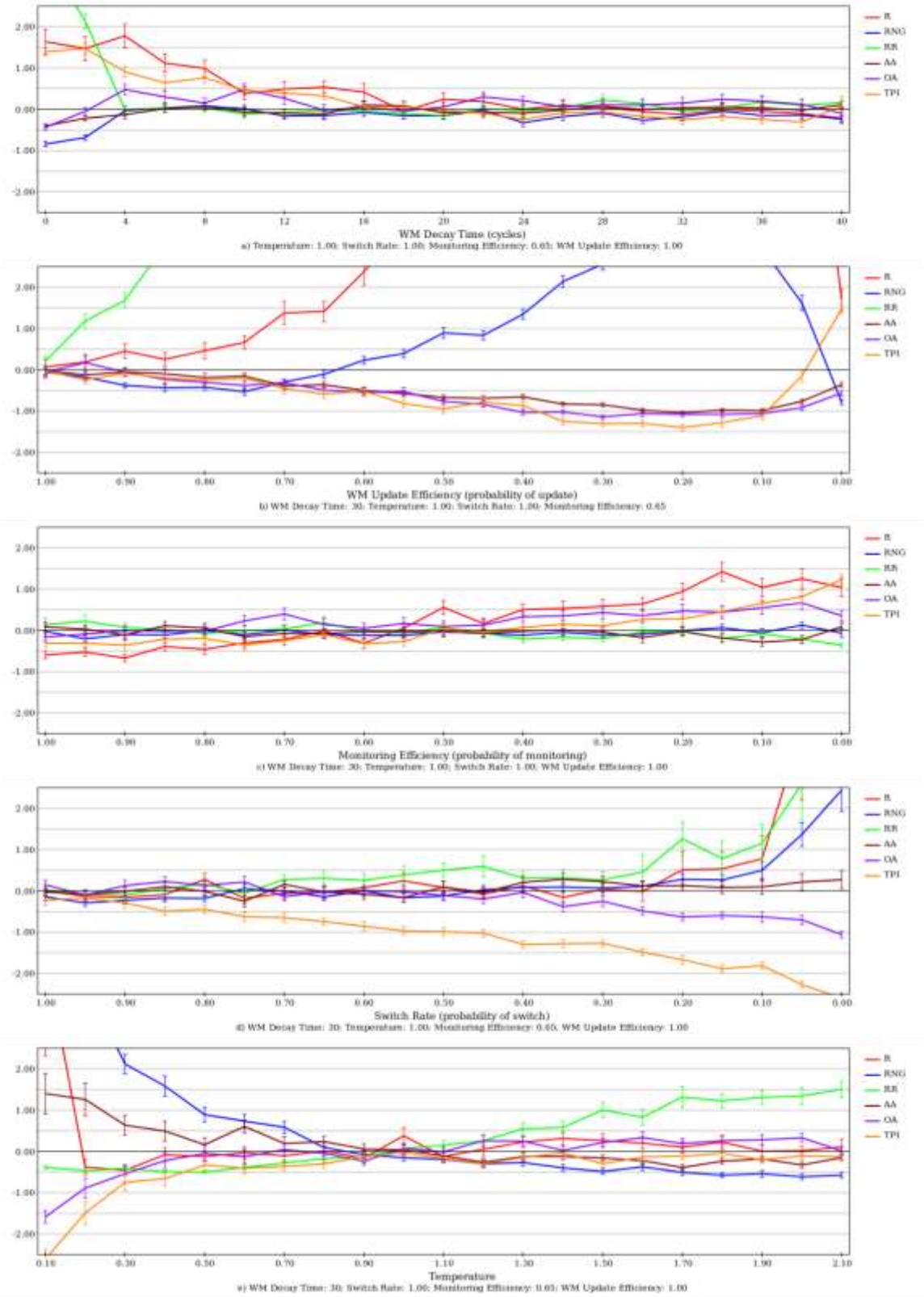

Figure S1: Effects of varying key parameters on each measure of randomness. Each panel shows the values of each measure of randomness as a function of the value of one parameter, with the measures of randomness in each case transformed to z-scores with respect to the data from the human subjects. This transformation a) allows all measures to be shown in a single graph and b) means that a value of zero corresponds to the group mean from the baseline subjects. Error bars represent one standard error.

The model's behaviour is less sensitive to the efficiency of the *Monitoring* process, as shown in figure S1 panel c). There is a broad range (from about 0.75 to 0.50) where all dependent measures are close to the values observed in the baseline subjects. Some measures (R and TPI) increase slightly when *Monitoring Efficiency* falls below 0.50, while R also decreases slightly when *Monitoring Efficiency* is greater than 0.75.

The effects of varying the parameters related to *Task Control* are shown in figure S1 panels d) and e). The primary effect of decreasing *Switch Rate* is a decrease in TPI, due to the model failing to switch (and hence failing to switch between clockwise and counter-clockwise selection) sufficiently frequently, though at very low values of *Switch Rate* other measures (most notably R, RNG, RR and OA) also diverge from the values obtained from human subjects. In contrast, variation in *Temperature* affects all measures of randomness. When the temperature is substantially less than 1.00, inherent schema selection biases (e.g., resulting in a tendency towards adjacent or opposite responses) are strong, while when the temperature is substantially greater than 1.00 these biases are weak, leading most notably to high rates of repeat responses.

7.3 Discussion

These simulations demonstrate that behaviour of the model is: a) relatively insensitive to *WM Update Rate*, provided that parameter is more than about 15; b) highly sensitive to *WM Update Efficiency*, which must be very near 1.0 for baseline performance; c) reasonably constant over a wide range of *Monitoring Efficiency*; d) dependent upon *Switch Rate*, which must be above approximately 0.8 or higher for baseline performance; and e) sensitive to the *Temperature* of *Task Control* which must be near 1.0 for baseline performance. As perhaps one might expect, the various dependent measures sometimes co-vary as parameter values are altered, but the dependencies between the dependent measures are complex.

While the simulations reported in this section do not rule out the possibility that the model might be tuned to account for any pattern of measures across the six dependent measures (cf. Roberts & Pashler, 2000), they do demonstrate the model's performance is not dependent on engineering the values of key parameters to specific values. With the exception of *WM Update Efficiency*, there is some freedom in the setting of parameter values needed to capture baseline performance. This is striking given that the simulations aim to fit the values of six dependent measures simultaneously.

Analysis of the model's behaviour when parameters are varied is also instructive. Miyake and colleagues (2000) extracted two principal components from the indices of randomness calculated from their subject data: equality of response usage, which was reflected primarily in subjects' R scores and held to reflect the efficacy of their *Memory Monitoring and Updating* executive function; and prepotent associates, which was reflected in TPI, RNG and adjacency measures and held to reflect the efficacy of their *Response Inhibition* executive function. Consistent with the former, in the current simulations, and with all other things being equal, very high values of *Monitoring Efficiency* result in lower values of R than more moderate values of *Monitoring Efficiency*, yet variation of this parameter (at the top end of its range) has no effect on other measures of randomness. The latter is less easily reconciled with the current model. However, prepotent associates tend to dominate responses when *Temperature* is substantially less than 1.0. Thus values of *Temperature* in the 0.7 to 0.3 range have a greater effect on several measures of prepotent associates (particularly RNG and AR). Whether this reflects ineffective inhibition processes or some other factor is unclear, though this issue is discussed further in relation to simulation 2 and in the General Discussion.