RUNNING HEAD: DISCRETIONARY TASK INTERLEAVING

Discretionary task interleaving: Heuristics for time allocation in cognitive foraging

Stephen J. Payne[a], Geoffrey B. Duggan[a] and Hansjörg Neth[b]

[a]University of Manchester

Manchester, UK

[b]Rensselaer Polytechnic Institute, USA

This manuscript has been accepted for publication within the Journal of Experimental Psychology: General. This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

Correspondence concerning this article should be sent to: Stephen Payne, Manchester Business School, University of Manchester, Manchester, M15 6PB, UK; Email: stephen.payne@manchester.ac.uk; Phone: 44(0)161 306 1280.

Abstract

When participants allocated time across two tasks (in which they generated as many words as possible from a fixed set of letters) they made frequent switches. This allowed them to allocate more time to the more productive task (i.e. the set of letters from which more words could be generated), even though times between the last word and the switch decision ("giving-up times") were higher in the less-productive task. These findings were reliable across two experiments using Scrabble tasks and one experiment using word-search puzzles. Switch decisions appeared relatively unaffected by the ease of the competing task, or by explicit information about tasks' potential gain. We proposed that switch decisions reflected a dual orientation to the experimental tasks. First, there was a sensitivity to continuous rate of return – an information foraging orientation which produced a tendency to switch in keeping with Green's rule (1984), and a tendency to stay longer in more rewarding tasks. Second, there was a tendency to switch tasks after subgoal completion. A model combining these tendencies predicted all the reliable effects in the experimental data.

Keywords

Task interleaving

Time allocation

Information foraging

Multi-tasking

Self-interruption

Discretionary task interleaving: Heuristics for time allocation in cognitive foraging

A fundamental problem for the human information processing system is which task to work on when. In the laboratory, especially in the problem solving laboratory, this dilemma is typically removed from the participant by design and instruction. But in everyday life, people typically have several active, independent goals and need to schedule their activities so as to allocate limited time adaptively across these goals.

Especially in today's high-pressure, high-information work environments, time management and multitasking have become important practical concerns. When workers have some discretion over what they do when, how might they schedule their activities, and how should they? If they are already working on one task, why might they give it up to work on another, only later to return to the first? For example, we suspect that many readers of this article have begun to write, but not yet completed, more than one research report. Why did they begin the second before finishing the first? And how do they decide when to work on which, when to pause writing one so as to work on another, and so on?

Experimental cognitive psychology does not have a great deal to say on these questions. In the experimental study of human performance, participants are only rarely given more than one task at a time, or any discretion concerning which tasks to do when. Nevertheless, cognitive psychology has investigated some important and fundamental limitations in people's capacity to multitask, and to swap between tasks.

Classic work on divided attention in the "dual task" paradigm (Styles, 1997) can, according to some theoretical treatments, be considered a matter of time-sharing across tasks, with mental resources being allocated to one or another task in turn. However, such

time sharing is at a very different level, in terms of timescale than the kind of task scheduling that is our concern in this paper.

More closely related is the popular recent experimental paradigm of task switching, in which performance is measured when people swap between two tasks, or perform the same task repeatedly. In a typical experiment in this paradigm, participants can respond to the same stimulus in one of two ways depending on the current "task set". Which task participants perform when is usually imposed by instruction, signaled perceptually, or both. For example, in the "alternating runs" paradigm (Rogers & Monsell, 1995) participants perform one task for two trials, and then the second task for the next two trials and so on. The fundamental finding is that there is a "switch cost". For example, in the alternating runs paradigm the first trial of each pair is performed slower than the second. Since Allport, Styles and Hsieh (1994), a large literature has grown around this topic. Theoretical interest has focused on the reasons for a switch cost, and the extent to which it reflects the operation of a unified central executive (Baddeley & Hitch, 1974).

Recent work by Arrington and Logan (2004) has moved one small step closer to our concerns by incorporating a voluntary aspect to the task-switch. Participants were asked to make either magnitude (greater than or less than 5) or parity (odd or even) judgments to digits. Unlike previous task-switching studies participants were allowed to choose which judgment to make of each stimulus, with instructions to balance the number overall and to produce a random order of judgments. As with the previous literature, task alternations were found to be slower than task repetitions, and this switch cost was higher at shorter RSIs. Participants' reasons for switching were presumably

nothing to do with task performance, instead being determined by the desire to balance tasks and to randomize order.

In summary, although this literature uses the name "task switching", the nature of the task demands and the central theoretical questions are very distinct from the focus of this paper, which is the voluntary, discretionary interleaving between independent tasks. Indeed, from our perspective, the basic finding in this literature, that task switching incurs a cognitive cost, makes mysterious our informal observation about everyday work, that people choose to interleave their activities. Why do people interleave activities, if every switch back and forth is slowing their task performance?

Applied work on multi-tasking has begun to address some of these issues. For example, a series of studies by Hockey and colleagues (Hockey, Wastell & Sauer, 1998; Sauer, Wastell, Hockey & Earle, 2003) investigated how people allocated effort across a set of tasks in simulated office or process control settings. In such studies the tasks were of different priorities (for one obvious example, consider driving, in which the task of safely reaching the destination is paramount, and tasks such as tuning the radio are clearly secondary; Cnossen, Meijman & Rothengatter, 2004). Interest has focused on how operators reflect this priority structure when adapting to task demands, such as fatigue or increased workload, with the main finding being that operators were able to successfully re-allocate effort so as to preferentially preserve performance on the higher priority tasks.

Of course there are many "extra-cognitive" environmental pressures that motivate switching between tasks in everyday life. Some interruptions demand immediate action. Some tasks take so long that they simply must be interleaved with other tasks that operate on different timescales. Nevertheless, our intuition is that people will often choose to

interleave activities when it is not strictly necessary, and it is this discretionary task switching that this article begins to investigate.

Some detailed studies of work activities in situ support our intuitions. Gonzalez and Mark (2005) studied multi-tasking in an office environment. They discovered a great deal of switching between "working spheres", much of it necessitated by external interruptions, but also a large number, as many as half of the switches, due to "self interruption" or discretionary switching. In both cases the workers were observed to prefer switching at "natural transitions", just after an action or sub-task was completed.

One can understand the prevalence of sphere-switches immediately after subtask completions by reference to the literature on imposed interruptions. A considerable body of experimental evidence has shown that interruptions have performance costs on resumed tasks (e.g., Hodgetts & Jones, 2006), but are less costly if they take place during subtask boundaries (Adamczyk & Bailey, 2004; McFarlane & Latorella, 2002). Thus choosing to switch immediately on subtask completion will be adaptive in terms of minimizing cost. However, paradoxically, if subtask completion were the only driver of switching behavior it would, in some situations, lead to maladaptive time allocation, because there would be a higher tendency to switch out of tasks with many subgoal successes.

Our basic experimental paradigm was straightforward. Participants were given a fixed amount of time to work on two tasks of equal priority that contributed entirely additively and independently to the overall goal. The first question we asked was simple – would people choose to switch between tasks in a situation like this? And if so, what would be the characteristics and causes of switching behavior?

There is some similarity between this paradigm and the six element test introduced by Shallice and Burgess (1991). In their test, participants were given six open-ended tasks to work on in a fixed period of time (15 minutes in the original version, with some constraints on which transitions are allowed). However, despite apparent similarities, interest in the six element task and its derivatives has focused on very different issues to those of concern in this article; in particular it has addressed frontal lobe involvement in abstract planning processes (e.g. Burgess, Alderman, Emslie, Evans, Wilson & Shallice, 1996) and the modeling of high-level planning processes in terms of a Supervisory Attentional System (SAS, Norman & Shallice, 1986; Cooper & Shallice, 2000). As far as we are aware, there has not been any emphasis on the details of the strategies that healthy participants use to solve the scheduling problem, over and above the abstract characterization of general processes. For example, granted that people multi-task by setting up and/or responding to environmental or psychological triggers (as proposed by SAS), exactly what triggers are attended to so as to switch from one task to another?

In the absence of any prior cognitive theories of discretionary task switching, we look to the animal literature for some theoretical guidance. In particular, we follow Pirolli and Card (1999), who, like others (e.g., Smith, Gilchrist & Hood, 2005, Sandstrom, 1994), have extended optimal foraging theory to human activities by comparing human behavioral solutions to solutions that optimize the rate of information gain, analogous with the way optimal foraging theory understands animal foraging by comparison with solutions that optimize rate of energy gain. In support of this general orientation, adaptive allocation of effort to information sources has recently been shown in the literature on

study-time allocation (Metcalfe, 2002) as well as in on-line browsing of expository texts (Reader & Payne, in press).

Optimal foraging theory is essentially an economic approach to foraging behavior. Animals are assumed to inhabit a patchy environment in which the energy gains in patches must be weighed against the costs of moving between and within patches (including obvious costs like the energy expended in locomotion and other tacit costs such as the risk of predation).

A major class of models, centered on Charnov's marginal value theorem (Charnov, 1976), considers the conditions under which an animal should leave one patch (a tree of berries, for example) to travel to another. The marginal value theorem itself states that the optimum solution to this problem is to leave a patch when the marginal rate of gain of energy is equal to the average rate of gain. However, this solution is not best viewed as a theory of what the animal actually computes, even if it achieves this solution. Going beyond analyses of what is optimum, researchers have considered the heuristics that animals might use to approximate this optimal solution in deciding to leave a patch. Most of this work has been concerned with visits to patches in sequence, rather than in alternation, with patches that are randomly encountered and with known reward functions. Nevertheless, at an abstract level the models offer plausible possibilities as descriptions of temporary deferral as well as of abandonment, and can certainly be applied to patches whose reward function is unknown. Of course, whether they are accurate models in this circumstance is an entirely empirical question.

Stephens and Krebs (1986) described four types of heuristic that might underlie patch-leaving decisions, or more accurately, perhaps, four variables that the decision may

rely on: time in patch; number of prey items encountered; "giving-up time" (time since last encounter of an item); and rate of encounter of items. The simplest heuristics would be to leave a patch after a certain period of time or after a certain number of encounters. It is easy to see, informally, that how well these simple rules perform depends on characteristics of the patches in question. For example: if some patches contain no prey items at all, then the number-of-items rule could be disastrous; even if this is not true, the rule will lead to richer patches being abandoned more quickly; if patches are very variable in the number of items they contain, then either rule is likely to be grossly inefficient.

A somewhat more sophisticated rule relies on the idea of "giving-up time". Giving-up time is the time between the last encounter of a prey item and the decision to leave the patch. The assumption here is that an animal sets a kind of patience threshold of acceptable duration between rewards and leaves once this is exceeded. A giving-up time rule performs well when the number of prey items per patch varies considerably (Iwasa, Higashi, & Yamamura, 1981).

More sophisticated again (at least in terms of its assumptions about what the animal computes) is a rule based on rate of encounter. One suggestion for how animals may be sensitive to rate of return across an entire visit to a patch is to assume that they track an estimate of the potential of a patch. If this estimate increases by a fixed amount with each encounter, but drops by a fixed amount per unit time (as in the fixed-time rule), then the potential will drop below threshold according to the rate of return across the visit. This mechanism is called "Green's assessment rule" by Stephens and Krebs (Green, 1984; Stephens & Krebs, 1986, p. 174), and fits with an analogy presented by

Green (1984) of a clockwork toy which winds down over time, but is wound up a little on every encountered item (Green's main presentation of the rule relied on dynamic programming to achieve a guessed rate, assuming it is the best possible). Similar mechanisms were proposed in earlier articles by McNamara (1982) and Waage (1979; based on empirical data from a parasitic wasp). Green's rule is explicitly targeted on patches where the reward function is not known. If one makes switch decisions according to Green's rule then completing a subtask will always postpone a switch decision rather than encourage it. Yet Green's rule would lead to overall adaptive time allocation in the case of patches of variable richness, because it leads to longer visits to richer patches.

To understand the behavior of any such rule in detail requires it to be parameterized (for example, to know over what time period the rate is computed). Nevertheless it is clear that the two more complex rules (based on leaving-time and rate), unlike the two simplest, will lead to an animal staying longer in richer patches.

The language of patches, encounters and prey items should not obscure the possibility of applying these general models to discretionary task switching by humans. In our experimental tasks, for example, participants tried to generate or find words. The analogy we make is between tasks and patches, and between the discovery of words and the encountering of prey items. A time-based leaving rule would see participants allocate roughly equal time to tasks, independently of their success at word generation. An item-based rule would see them persist with each task for however long it takes to generate roughly the same number of words. A rule based on a giving-up time threshold, or on a rate-of-return threshold would appear to be more adaptive, in that it would lead to

participants spending greater amounts of time on tasks that were producing more frequent rewards.

Analogous issues about choice-rules used by animals have arisen in the literature on operant conditioning, and especially work on the phenomenon of "matching". In a classic matching experiment an animal is free to respond at two different levers, each programmed with a different reward schedule. Over time, the ratio of response to reward at the two levers is observed to be approximately equal, which empirical regularity is called "matching". There are several theories of how and why matching is achieved, for a review see Miller and Grace (2003). Depending on the particular reinforcement schedules, it is sometimes, but not always the case that matching results from optimal response allocation (e.g. Herrnstein & Heyman, 1979). In such cases, one proposed mechanism of how matching could be achieved is melioration, in which the animal is assumed to respond to whichever lever has the currently higher level of local reward. With the schedules typically employed, this heuristic requires fairly frequent switching of activity between levers. Because melioration is closely related to hill-climbing, which is a well-established strategy in human problem solving (Newell & Simon, 1972), this analysis suggests a rather obvious link between these operant experiments and our experiments, which will be explored once the basics of our experimental set-up have been described.

To investigate whether humans use heuristics of this sort to make task switching decisions, we needed an experimental situation in which participants were given a set of independent tasks, each of which contributed to their overall performance, and could switch freely between tasks, choosing when to work on each task.

The work on foraging theory shows that which rules of thumb are most successful depends critically on the nature of the patches, and in particular on the distribution of rewards (Stephens & Krebs, 1986). Particularly important is the expected gain curve – the distribution of rewards over time on a task. In the operant literature these gain curves are determined by ratio or interval schedules, and the "patches" are non-depleting, but intuitively these do not seem typical characteristics of human tasks or activities. In general we imagine that human tasks might have a variety of gain curves (where gain is measured in terms of objective returns, or subjective ones like satisfaction), and that the shape of these curves might be an interesting way of classifying tasks and understanding aspects of behavior, including task switching. However, for our explorations of this topic, we decided to focus on tasks with a gain curve that we believe to be typical of a great number of human activities that are extended in time (and indeed to many foraging situations), namely a monotonic pattern of diminishing returns, in which the gain increases with time on task, but gain per unit time drops gradually until eventually the task is effectively depleted (as an everyday example consider revising a manuscript).

In summary, we have briefly reviewed a large number of disparate literatures that have some relevance to our concern. Despite superficial similarities, we argued that the human experimental literatures on task switching and executive control are only weakly related to our goals. To understand whether and how people interleave activities we have combined ideas from two fields. From foraging theory we took the perspective that switch decisions may be determined by characteristics of activities in terms of the occurrence of successes over the period of an activity. From applied work on multi-tasking we took the perspective that agents may seek to minimize switch costs by

managing transitions according to sub-task boundaries. These perspectives make different predictions: the foraging rules (with the exception of the fixed item rule) suggest that giving-up decisions will tend to be remote from task completions, whereas the cost-management perspective suggests the opposite. We explored these issues within a simple experimental context: this allowed us to uncover new empirical regularities concerning the effectiveness of overall patterns of time allocation across tasks as well as fine-grained temporal measures concerning task-switch decisions.

Experiment 1

The first experiment examined baseline performance on two of the "Scrabble" tasks that were later to be used to study time allocation between tasks. (Our participants' task was a degenerate version of the well-known game Scrabble, in which their goal was to make as many words as possible from a set of letters.)

There were several criteria determining the choice of task. First, there should be a ready index of performance or gain that was available to experimenters and also to participants (so that choices to switch task could be considered in terms of overall and moment –by-moment performance). Second, cumulative performance indices should increase continuously and monotonically with time spent on task. Standard experimental problems with a single goal, such as the Tower of Hanoi, were inappropriate for this reason. In addition to requiring a task in which measures of performance increased over time, we preferred a task in which this function was one of diminishing returns, so that gain per unit time decreased.

It was important that switching between tasks could in principle be entirely discretionary, and not logically necessary for performance, so that there should be no

dependency relations among the set of tasks. Furthermore, it was essential that there should be a valid indication of which task was being worked on when. The Scrabble task (making words from a set of letters) had the advantage of being very dependent on a visual stimulus (the array of letters), making it at least very likely that participants would choose to display the set of letters before and during their attempts to generate words from that set (contrast this with, for example, a category fluency task of Maylor, Chater & Jones, 2001 - in which case participants may have generated exemplars from a non-current category before indicating their switch.)

Method

Participants

Participants were 50 undergraduates of Cardiff University, 16 male, 34 female, ages from 18 to 23, who took part for course credit.

Design

Each participant was assigned to one of two Scrabble tasks, defined by the set of 7 letters from which the participants were instructed to generate as many words (of at least two letters in length) as possible.

Materials

The two sets of seven letters from which participants were required to make words were taken from Maglio, Matlock, Raphaely, Chernicky and Kirsh (1999). These authors presented randomly generated sets of 7 letters to participants for 5 minutes and required them to construct as many words as possible by rearranging subsets of the letters. Maglio et al. (1999) reported the highest mean number of words generated was 26.1 for the set “LNAOIET”, so this set was used for the “Easy” task. The lowest mean

number of words generated was 12.1 for the set “ESIFLCE” and so this was used for the “Hard” task. We further checked the potential yield of the sets of letters, using a program called “scrabble buddy” (http://boulter.com/scrabble//). Scrabble buddy generated 154 words for the Easy set, and 64 for the Hard, but this included proper nouns and acronyms, and, furthermore, not all of these words were recognized by us. We presented each set of words to a single student participant and asked him to check which words he recognized of those that met our constraints. This process yielded 53 words from the Easy set and 23 from the Hard set. We regard these figures as good estimates of theoretical maximum performance at each task.

A program was written in MS Visual Basic 6.0 to present appropriate letter sequence and record participants’ performance. Below the letters was an Entry box that displayed the letters of a word as they were typed by the participant and a button labeled “Enter”. In the top left of the screen was a button labeled “Start” and beneath this was a timer box that displayed the number of seconds remaining to generate words. Each time the Enter button was clicked on using the mouse, the program recorded the contents of the Entry box and time stamped the event.

Procedure

Participants were instructed to generate words using the letters provided. Participants were informed that words did not have to use all the letters but had to be between 2 and 7 letters in length, and that proper nouns or acronyms were not allowed.

After clicking on the Start button a clock displayed the number of seconds remaining and began to count down to zero. Participants could type words into the Entry box using the keyboard. After each word had been typed participants were required to

click on the "Enter" button, this cleared the Entry box in preparation for the next response. After 600 seconds had elapsed the experiment was automatically terminated. The program did not provide any feedback about any errors in the words entered.

Results and Discussion

Words generated by participants were counted by hand. Any answers that used the letters provided and were listed on www.dictionary.com were counted as words. In practice this was uncontroversial (i.e., there were no very rare actual words that people produced). Occasionally, participants would erroneously use a letter that occurred only once in the set of letters more than once in a word, and such answers were of course discounted. Words were also occasionally repeated, and only first occurrences were counted in the totals.

On the Easy task participants generated on average 29.77 (SD = 8.10) words, and on the Hard task, 14.55 (SD = 5.10) words. The cumulative word-generation graphs over time showed a classical diminishing returns shape (see Figure 1). After five minutes, participants doing the Easy task had generated on average 21 words, and participants doing the Hard task had generated 11.3 words. Using the average data shown in Figure 1, we can estimate the optimal time allocation across the tasks given an overall budget of 10 minutes. Figure 2 shows that the optimal overall time allocation is to spend approximately 25% of time on the hard task.

It is interesting to examine how this empirical analysis relates to the matching hypothesis. Spending 2.5 minutes on the hard task, led to an overall rate of return on that task of 3.51 items per minute, and an overall rate of return of 3.49 items per minute for 7.5 minutes on the easier task, for the averaged data. Thus optimal performance

approximated the matching principle. One simple mathematical model for approximately fitting the diminishing returns curves of Figure 1 is the exponential function, Figure 1 shows two such functions. Dividing the 10 minutes into 100 units (U) of 6 seconds, the functions we used (and display as 'model' in Figure 1) were:

Hard Task: Words per unit = $.51 \times .965^{U-1}$; Easy Task: Words per unit = $.65 \times .981^{U-1}$

Experiment 2

Having quantitatively characterized serial performance on two Scrabble tasks of varying difficulty, we investigated performance when the two tasks could be interleaved. Although we were primarily interested in a condition where each participant could allocate a single budget of time freely between two tasks, we also considered a condition in which participants could switch between tasks at will, but only until they had used a fixed budget of time (5 minutes) for each task. This comparison between "Interleave-Free" and "Interleave-Equal" allowed us to investigate the hypothesis that switching between tasks was a strategy for allocating time preferentially. This hypothesis would be supported if the Interleave-Free condition elected to switch tasks more often than the Interleave-Equal condition. We additionally considered conditions in which participants worked on one task then the other, in either order. This "serial" condition allowed us to assess the magnitude of switch costs (by comparison with the Interleave-Equal condition).

Method

Participants

Participants were 72 undergraduate students (14 male, 58 female) from Cardiff University. Their age ranged from 18 to 39. The participants were each paid £3 [Footnote 1 here] or given course credit in exchange for completing the experiment.

Design

The freedom that participants had to allocate their time between the two tasks was manipulated between participants. Participants were either allowed to switch freely between the two tasks or were required to complete each task sequentially in a serial fashion (the serial condition constituted a within-subjects replication, for five minutes only, of Experiment 1 and allowed direct comparisons of performance). The total time spent on the two tasks either had to be divided equally between the tasks or was determined by the participant (equal or free). The combination of these variables produced three conditions labeled Interleave-Equal, Interleave-Free and Serial-Equal.

Each participant received the same "Easy" and "Hard" tasks used in Experiment 1. Overall performance was measured in terms of the total number of distinct words generated, and participants were informed of this criterion. The time of entry of every word, and the time whenever a participant switched tasks were recorded.

Materials

As before a program was written in MS Visual Basic 6.0 to present the two letter sequences and record participants' performance on both tasks. Two buttons labeled "Sequence 1" and "Sequence 2" were horizontally aligned towards the top of the screen. Otherwise the screen components and their position were as in Experiment 1. In the Interleave-Equal condition there were two clocks, adjacent to the selection buttons for each task; each displayed the seconds remaining for that task. In the Serial-Equal

condition a single larger button labeled “Sequence” replaced the two buttons labeled “Sequence 1” and “Sequence 2”. These were the only differences among the interfaces for the three conditions.

Each time either of the Sequence buttons or the Enter button was clicked on using the mouse the program recorded the contents of the Entry box, the selected task and time stamped the event.

Procedure

Participants were instructed to generate words from the letters provided, as in Experiment 1. In the interleave conditions participants were informed they could switch between the two letter sets as much or as little as they wanted, but their aim should be to maximize the total number of words generated.

After clicking on the Start button the clock or clocks displayed the number of seconds remaining and began to count down to zero. When participants clicked on a Sequence button the corresponding sequence was displayed thereafter in the box above the Entry box and participants could then type words into the Entry box using the keyboard. After each word had been typed they were required to click on the “Enter” button, this cleared the Entry box in preparation for the next response. In the interleave conditions participants could change the sequence displayed at any time by clicking on the corresponding Sequence button. After 600 seconds had elapsed the experiment was terminated.

In the Serial-Equal condition the clock displayed 300 seconds at the start of a task and when this time had elapsed the whole procedure was repeated immediately for the other letter sequence. In the Interleave-Equal condition both clocks initially displayed

300 seconds. The clock on the left side of the screen corresponded to the sequence button on the left side of the screen and vice versa. Each clock remained static until the related sequence was selected and then it began to count down. Once either of the sequence buttons had been clicked on for the first time one or other of the sequences was always selected. This meant that a clock was always counting down from that point onwards. When either clock reached zero both the corresponding sequence button and the Entry box were disabled. Clicking on the button for the uncompleted sequence reactivated the Entry box and the corresponding clock. In the Interleave-Free condition the single clock simply displayed 600 seconds at the start of the experiment.

Presentation of the sequences was counterbalanced so that they appeared an equal number of times both on the left or right of the screen in the interleave conditions and first or second in the Serial-Equal condition. The program did not provide any feedback about any errors in the words entered.

Results

Correct responses were identified and counted as before. However, all attempts, whether "correct" or "incorrect" were treated as items in the timing data. (6% of responses were categorized as incorrect.)

First, consider the number of words produced in the three main conditions (pooling over the counterbalancing of order in the design). Table 1 shows that, as expected, participants in all three groups generated considerably more words from the "Easy" letter set. A 3 × 2 (Condition × easier/harder task) mixed ANOVA revealed a significant main effect of this difference, $F(1, 69) = 363.00$, $MSE = 14.78$, $p < .001$, $\eta^2_p =$

.84. However, neither the difference in the number of words generated across conditions nor the interaction between Task Difficulty and Condition were significant ($\underline{F}$s < 3).

It is somewhat surprising that the Interleave-Free group did not manage to generate more words from the Easy sequence, as 23 of 24 participants spent longer on the Easy sequence, (binomial test, $\underline{p} < .001$; ANOVA was inappropriate for the time allocation data because time on one task was a residual of time on the other). On average Interleave-Free participants spent 59% of their time on the Easy sequence and generated 73% of their words from this sequence. Thus, although participants were successfully managing to devote more time to the easier task, they were "under-matching", in that they were spending more time on the Hard sequence and (according to the data from Experiment 1), less time on this task than would be optimal.

Participants in the Interleave-Free condition switched tasks on average nearly 7 times in 10 minutes. One of the reasons for this appeared to be in order to allocate time preferentially, as participants in the Interleave-Equal condition switched significantly fewer times, $\underline{t}(46) = 2.11$, $\underline{p} < .05$, $\underline{\eta}^2_p = .09$. However, the Interleave-Equal participants still showed a substantial tendency to switch.

We now look in detail at the timing protocols to test for the various heuristics that may have driven giving-up decisions. We use the term "visit" to denote a period of working on one of the tasks; each visit ended when the participant switches to the other task. Giving-up times (time between a click on Enter and a click on the non-current Sequence button) and longest between-item times (the longest times between two consecutive words (Enter presses) on each visit, averaged over visits then participants) are displayed in Figure 3. Data from one participant in the Interleave-Free condition and

two participants in the Interleave-Equal conditions were excluded from Figure 3 and subsequent analyses as they made only one switch during the experiment. For the remaining data, means were taken for each participant, ignoring the last visit (which was ended by the timer rather than by the participant's active choice) and visits in which no items were generated, and then averaged across participants. Analyses were done separately for the Interleave-Free and Interleave-Equal groups, as we were primarily interested in the decision behavior of the Free group.

Figure 3 shows that giving-up times were longer in the Hard task. For the interleave-free condition this difference was significant, $t(22) = 2.69$, $p < .05$, $\eta^2_p = .25$. This trend was replicated in the Interleave-Equal condition but was not significant, $t(21) = 1.61$, $\eta^2_p = .11$.

Figure 3 also shows that across all participants across both tasks the average longest between-item time was greater than the average giving-up time in both the Interleave-Free condition, $t(22) = 4.86$, $p < .001$, $\eta^2_p = .52$, and the Interleave-Equal condition, $t(21) = 3.10$, $p <. 01$, $\eta^2_p = .31$.

The mere fact that participants spent reliably more time on the easy task rules out the two simplest giving-up rules, based on time or on number of items generated. The time rule predicts no reliable difference between the tasks, whereas the number rule predicts that participants would spend more time on the more difficult task.

A simple giving-up-time rule would predict the observed tendency to spend longer on the easier task. However, such a rule could obviously NOT explain the reliable difference in giving-up times between tasks. Further, as noted above, we discovered that longest between-item times tended to be longer than giving-up times.

Is it possible that participants were using a rate-based leaving heuristic like Green's assessment rule (Green, 1984)? Green's rule appropriately predicts longer overall times on easier tasks. Less obviously, it also predicts longer giving-up times in the harder task than the easier task. This was an unexpected but reliable aspect of the data.

Consequently, Green's rule apparently simultaneously predicts two reliable effects, that appear to work against each other – longer visits to the Easy task, but shorter giving-up times for this task. Before modeling these two effects quantitatively, we will attempt to explain why Green's rule predicts the effects.

According to Green's rule the length of a visit is determined by the equation $V = T + IG$, where V = Visit time, T and G are free parameters, and I is the number of Items generated during the visit. T can be considered as the minimum visit time, i.e., the Time if no items are generated, G is the Gain in visit time for each item generated.

This equation straightforwardly predicts longer visit times to the easier task, because more items were generated on this task.

The prediction for giving-up times is harder to intuit. Although Green's rule leads to longer visits to Easier patches, it also leads to patch visits with higher rates of return when the patch is richer – i.e., it doesn't lead to an increase in visit times that is sufficient to equalize rate of return. The rate of return during a patch visit is $I / (T + IG)$, and thus increases with I. As a consequence, we can see that Green's rule will tend to produce smaller giving-up times in richer patches. In richer patches the visits were longer, but items occurred more densely within that time period. Therefore the chances of an item occurring shortly before the leave decision (producing a short giving-up time) are

increased. Shorter giving-up times in richer patches (easier tasks) thus occur as a probabilistic effect.

This argument will be validated by the quantitative model we are about to describe. However, before we describe the model we want to anticipate the conclusions of the modeling enterprise: we discovered that it was possible to fit the qualitative main effects on visit time and giving-up time with Green's rule, but not the size of the effects. It proved impossible within the constraints of our model to reproduce the size of the difference in giving-up times at the same time as size of the preference for the Easy task.

The model generated a "run" through the experimental task. The 600 second experimental time was divided into 100 units of 6 seconds [Footnote 2 here]. Within each unit of time, an item was generated with a probability according to the functions we fit to the participant data in Experiment 1. Thus initial probabilities of generating a word in a 6 second window are .65 for the Easy task and .51 for the Hard task. These probabilities were multiplied by .981 (Easy) and .965 (Hard) in each time increment.

Also in each unit of time a decision was made about whether or not to switch task, using Green's rule – which we implemented so as to model the graphical description in Stephens and Krebs (1986, p. 175), and will describe, following them, using the idea of a patch having a potential which decreases over time and increases when items are found. As soon as a task was begun (either at the start of the experiment or after a switch), the potential of the patch (the time that would be spent there if no items were found) was set to T (the minimum time parameter, as above). In each unit of time, potential was reduced by 6 seconds. If an item was generated, potential was increased by G seconds. When potential reached zero the task was switched.

A single run of the model, with the same parameter values, would produce variable time profiles, because of the probabilistic implementation of item generation. Thus, we tested the model by setting the two parameters and then running the model 200 times and using the resulting statistics.

At the start of a run of the model one of the two tasks was randomly chosen as the active task. The first unit of time was then executed using the active task to determine the probability of generation following the principles above. After the decision to switch task or not had been made the second iteration was then carried out in much the same way. When a decision to switch was made the active task was changed for the next unit of time. Generation was postponed within an inactive task so that the generation rate of the active task reflected the number of iterations completed within that task rather than the sum of both tasks. A run was complete when all 100 iterations had been carried out.

Following Roberts and Pashler (2000), Roberts and Sternberg (1993), we tested the predictions of this model for time on easy task, number of switches and giving-up times across a range of plausible values of G and T, according to our judgment. (The participant data for all variables was within the model's predictions, but these predictions were rather wide in the case of time on easy task and number of switches; a report of this phase of model-testing is available from the authors).

To provide a more stringent test of the model's ability to account for the data we fixed the number of switches using the participants' average data and varied the range of parameters to see if the model could simultaneously fit the 2 remaining dependent variables of primary interest: the proportion of time on the Easy task and the difference between giving-up times on the Hard and Easy tasks. (In effect, this tested whether the

model can simultaneously predict the quantitative values of 4 dependent variables: number of switches, time on easy task, giving-up time on easy task, giving-up time on hard task (these last two reduced to a single difference measure)).

To try to achieve the best possible fit, a fine-grain exploration of parameter values in the neighboring range was required. G values increased from 3 seconds in 3-second units. Then T values were varied in six-second units so as to find combinations of parameters that fitted the number of switches to 7 ± 1.01 (the participant mean and 95% confidence interval) and did not generate more than 5% highly discrepant runs of 0 or 1 switch. These best-fitting points are shown in Figure 4. Also included is the mean participant data from Experiment 2; around all data points a 95% confidence interval has been plotted using the variance from Experiment 2. Figure 4 indicates that the model can only account for about half the area within a 95% confidence interval around the mean from Experiment 2. It seems that Green's rule, although it can offer a plausible account of the qualitative patterns in the data, cannot fit the quantitative aspects of all our dependent variables simultaneously.

Another prediction that arises from Green's rule is that giving-up time must never be less than G (assuming at least one item is generated). Yet our exploration of the parameter space has indicated that when there are 7 switches G must be at least 6 seconds for the proportion of time spent on the Easy task in the model to approach the participant's mean in Experiment 2 (i.e. 59%). However, in Experiment 2, 26% (18/70) of the giving-up times from the Easy task and 16% (11/70) of the giving-up times from the Hard task were below 6 seconds. This suggests that Green's rule cannot explain all of the participants' decisions to switch task.

This issue is strongly aligned with our discussion in the introduction, concerning task switches at subgoal boundaries. We pointed out that in situ studies show that people sometimes switch spheres of activity immediately on completing a subgoal, and that this is a good strategy for minimizing the costs of self-interruption. Perhaps in our experimental task, participants sometimes adopted a subgoal orientation and chose to switch "immediately" on finding a word. The data we have just described support this contention.

Consequently, we explored a model in which switch decisions have two independent bases. One, based on Green's rule, arises from treating word-finding as an activity, and monitoring the changing gain curve of that activity. The second, based on subgoal completion, treats the current search for a single word as a subgoal and switches on its completion. To model the mixing of these two strategies we simply added one additional free parameter - the probability of switching on subgoal completion (finding a word). If this probability is low, then even in longer visits only a proportion of all task switches would be due to the subgoal strategy, the remainder being determined by Green's rule. This simple addition to Green's rule is to some extent determined by the data – it is a simple reflection of the observed minority of switches that occur very shortly after a word finding. However, it is also justifiable in terms of naturalistic observation, as noted above.

What does this model predict, across its combinatorial parameter range? The experimental participants switched on average 7 times and generated about 28 items. Thus, if the p-subgoal parameter was 0.25 it could account for all of the switches. We thus varied p-subgoal from 0.025 (0 would be the vanilla Green's rule case, as above) to

0.125 (although the rapid-switch data suggests that many fewer than half of the switches were "immediate"). In runs of the model, the relation between the p-subgoal parameter and the proportion of fast switches is not entirely straightforward; nevertheless this argument allowed us to set a plausible range for the parameter.

The final decision was how to assign a time to "immediate" switches within the constraints of our model - to prevent these switches having an unrealistically short giving-up time of 0 seconds, a lag was introduced so the switch occurred one unit of time (6 seconds) after the item was generated.

When the same constraints on parameters and number of switches were applied as for the Green's rule simulations above, the relation between giving-up time Hard-Easy and proportion of time on Easy task was as shown in Figure 5. Thus, the model predicts that behavior should be limited to a particular area of the plausible space of the dependent variables we are considering. Further, the participants' data was within this space. Thus the model makes a genuine prediction that is supported by the data. Further, although for ease of visualization we have considered the difference in giving-up times hard-easy, the model in fact predicts the appropriate absolute values for both these variables.

Finally, we come to the more conventional question about a quantitative model – could it fit the participants' data with judicious selection of parameter values. The graph shows that it can – the closest model point to the participant data in Figure 5 is achieved with these parameter values: T= 30 seconds; G= 18 seconds; p-subgoal = 0.1.

There is one dependent variable that we have not considered so far in our exploration of the quantitative models, namely the longest between-item times. With the above parameter values the model generates longest between-item times which

underestimate the participants' average data (we have not pursued this discrepancy: participants' longest between-item times are essentially outliers in their performance profiles, and therefore even when averaged may prove hard to fit with a model targeted on explaining average performance). Thus our quantitative model fits with the simultaneously observed values of 4 of 6 dependent variables (all within the 95% confidence intervals of the means), for values, see Appendix 1.

Discussion

This experiment discovered several intriguing phenomena. First, people, in general appeared to be inclined to switch between tasks, despite the evidence from comparisons with the Serial-Equal condition that this had no overall benefit for performance. Part of the reason for so doing was presumably to allocate time adaptively – most of the participants spent more time on the Easy task, and by so doing should have improved their total score relative to some alternative schedules (e.g. all the time on one task). However, according to the principle of matching, participants would have further benefited by allocating even more of their time to the easier task. The fact that participants in the Interleave-Equal condition also switched between tasks, although at a reliably lower rate, suggests that time-allocation was not the only motivation for switching.

The local determinants of giving-up decisions were not straightforward. The combination of three reliable effects is a challenge for models of switch-choice. People spent more time on the Easy task, but giving-up times were typically lower than the longest between-item time in a task-visit, and giving-up times were reliably, and substantially higher on the hard task.

In combination, these findings ruled out the most straightforward models of switching based on thresholds of time, items, giving-up time, or rate of return. Green's rule offered an attractive possibility, because it naturally explains the successful allocation of more time to the richer task, as well as producing a difference in giving-up times in the reported direction. However, our attempts at quantitative modeling showed that it cannot simultaneously fit the size of these effects.

We proposed a dual-process model, inspired by the observation of a substantial minority of very fast giving-up times. According to this model, Green's rule is elaborated with a probabilistic component that chooses to switch immediately on completing a subgoal (generating a word). We modeled this component as a simple independent probability of switching on subgoal completion. This combined model captures the informal distinction made in the introduction between stopping rules based on ongoing activity, monitoring a continuous measure of gain and using this as the basis for any switch (as in foraging generally), and orientation to a single subgoal, with stopping determined by subgoal completion (as in conventional human problem solving tasks). The model only has three free parameters. Varying these combinatorially across their plausible range showed that the model makes genuinely constrained predictions, and the participants' data supported these predictions. The model allowed good quantitative fits of the primary dependent variables (participant means) as well as providing an explanation for all important qualitative effects. (Although it underestimates longest between-item times, it correctly predicts that longest between-item times were longer than giving-up times for the Easy task – c.f. Figure 3.)

Nevertheless, a quite different possibility that seems particularly salient in the context of switching back and forth in a small set of tasks (especially two tasks) is that switch decisions were based on the characteristics of alternative tasks as well as the current task. This is the fundamental assumption of models of hill-climbing or melioration (Herrnstein & Vaughan, 1980) in the operant conditioning, matching literature.

Experiment 3

The second experiment was designed to test the hill-climbing heuristic, as well as to test our dual-process model by replicating the important reliable findings of the first experiment with regard to the switching behavior of the Interleave-Free group. A third set of letters was introduced, of intermediate difficulty. One group of participants was presented with this set and the Easy set from Experiment 2, a second group of participants was presented with the Medium set and the Hard set from Experiment 2. Both groups were free to switch between tasks as they preferred, and our interest focused on switching behavior, and, in particular, on whether decisions to switch from the Medium task were influenced by whether the alternative task is harder or easier.

Method

Participants

Forty Cardiff University undergraduate students (15 male, 25 female) were each paid £3 in return for participating in this experiment. Their ages ranged from 19 to 33.

Design

All participants completed the tasks under the Interleave-Free condition from the previous experiment. However, one of the two tasks to be completed was varied between

participants. Each participant carried out the same task of medium difficulty. However in conjunction with this task half the participants had to carry out an easy task (the Medium/Easy group) and half the participants had to carry out a hard task (the Medium/Hard group).

Materials

The Easy and Hard tasks used the same letter sequences from Experiment 1. The letter sequence for the Medium difficulty task was also taken from Maglio et al. (1999). The sequence "NDRBEOE" was chosen as the closest to midway between the easy and difficult tasks from their data. (They reported a mean of 19.88 words generated in 5 minutes.) The words generated by scrabble buddy for this sequence were filtered by our single student participant, as for the Easy and Hard sets, and produced 35 words as a best estimate of maximum performance from this set of letters.

Procedure

The total amount of time participants were given to generate words was increased from 600 seconds to 840 seconds. This enabled us to investigate switching behavior over a longer period of time, with (presumably) more visits contributing to the mean data on visit-times, giving-up times and so on. All other aspects of the methodology were identical to the Interleave-Free condition from Experiment 2.

Results

Table 2 shows the performance measures for the two groups, divided by task/ letter set. Participants generated more words in their easier task rather than their harder task (4% of responses in the Medium/Easy and 7% of responses in the Medium/Hard group were categorized as incorrect and only treated as items in the timing data). A 2 × 2

mixed ANOVA (Group, Easy/Medium v Easy/Hard × Task Difficulty, easier or harder) showed a main effect of task difficulty, $F(1, 38) = 54.97$, $MSE = 27.86$, $p < .001$, $\eta^2_p = .59$ – participants generated more words on their easier task, and a main effect of Group – participants who received the Medium/Easy tasks generated more words in total than those who received the Medium/Hard tasks, $F(1, 38) = 10.80$, $MSE = 58.10$, $p < .01$, $\eta^2_p = .22$. However, the interaction between these two factors was also reliable, $F(1, 38) = 9.31$, $MSE = 27.86$, $p < .01$, $\eta^2_p = .20$, with the difference between tasks being higher in the Medium/Hard group.

As in Experiment 2, participants in both groups allocated more of their time to the easier task, (overall, 28 out of 40 participants; binomial test, $p < .05$), but not to the extent that would be required to match the cumulative rates of return. This effect is noticeably smaller than in Experiment 2, perhaps unsurprisingly given that the difference in difficulty between any participant's pair of tasks had been deliberately reduced.

Turning to the timing data for giving-up decisions, data from two participants in the Medium/Easy group and one participant in the Medium/Hard group were excluded as they made only one switch during the experiment. Figure 6 shows that, again as in Experiment 2, giving-up times were shorter than the longest between-item times. A 2 × 2 ANOVA treating the time measured per visit as a within-subjects variable (Group × Time Type, Longest vs. Giving-up) found this main effect was reliable $F(1, 35) = 36.72$, $MSE = 42.23$, $p < .001$, $\eta^2_p = .51$. Additionally, there was a main effect of Group: across both measures the Medium/Hard group took longer than the Medium/Easy group, $F(1, 35) = 4.87$, $MSE = 170.90$, $p < .05$, $\eta^2_p = .12$.

Because our patch-leaving hypotheses make stronger predictions for giving-up times, these times were analyzed separately, including easier/harder task as a within-subjects factor. There was a main effect of this factor: giving-up times were longer in whichever was the more difficult task of the pair, $F(1, 35) = 10.84$, $MSE = 183.92$, $p < .01$, $\eta^2_p = .24$, as well as a main effect of Group, with giving-up times longer in the Medium/Hard group than the Medium/Easy group, $F(1, 35) = 6.90$, $MSE = 211.26$, $p < .05$, $\eta^2_p = .17$. The interaction between Task Difficulty and Group was also significant, $F(1, 35) = 4.43$, $MSE = 183.92$, $p < .05$, $\eta^2_p = .11$.

Thus, all three of the most important findings from Experiment 2 were replicated: participants spent longer in their easier task, and had longer giving-up times in their harder task. Giving-up times were shorter than longest between-item times within a visit.

We now turn to the question of hill-climbing, i.e., the role of the competing task on giving-up decisions. If decisions to leave a task were influenced by the competing task, we would expect visits to the Medium task to be shorter when it was paired with the Easy task. Table 2 shows a small difference in this direction, but it does not approach significance $t(35) = .57$, $\eta^2_p = .007$. (Participants' last visits were excluded from this analysis, as they were timed out.)

One might argue that this test is low power, in that average visit times were rather variable because of the variation in overall number of switches. For this reason, we inspected each participant's first visit to the Medium task when they had experienced the competing task (i.e., either their second or third visit overall). These data showed no effect of competing task – visits to the Medium task were slightly but not reliably longer on average when the competing task was Easy, $t(35) = .27$, $\eta^2_p = .002$.

A very similar test can be made of the influence of the difficulty of the current task (as opposed to competing task), by inspecting each participant's very first visit. At this point the participant has absolutely no knowledge about the competing task, yet in the Medium/Hard group the length of the first visit to the easier task was shorter than to the harder task, $t(18) = 3.04$, $p < .05$, $\eta^2_p = .34$. This difference was not significant in the Medium/Easy group $t(18) = .13$, $\eta^2_p = .001$, but the performance data showed that these two tasks were not very different in difficulty.

A similar story emerged from the giving-up times. If the main influence on these times were the attraction of the competing task, then we would expect giving-up times for the Medium task to be shorter when the alternative is Easy rather than Hard. There was no such tendency in the data (see Table 2).

Having found no evidence for the plausible hill-climbing account of switching, and replicated all the main qualitative effects of Experiment 2, we turned to the question of whether our quantitative model of switching can explain the current data. Having already shown, in dealing with the results of Experiment 2, that the model makes seriously constrained predictions, we focused on whether the model can fit the new data, using the old best-fitting parameter values, scaled to the different time dimensions of this experiment. Total time available to participants increased from 10 to 14 minutes so we increased T accordingly from 30s to 42s. Because G 's role is to allocate time preferentially to better patches it is unclear whether or not it should be scaled, so we explored two parameter-sets, one in which G was scaled (24s instead of 18s) the other in which G remained the same. The value of p-subgoal was held constant at 0.1. (N.b. the initial word generation probabilities for the medium task were chosen to best fit the

overall word generation data, yielding slightly different values in the two conditions, .57 when paired with Hard and .50 when paired with Easy.) The model was extended to 140 time units each corresponding to 6 seconds.

As shown in Appendix 1, the resulting model produced good fits of all dependent variables. Predictions for all six variables were within the 95% confidence intervals of participant means. The underestimate of longest between-item times was repeated but ameliorated, and, again all important qualitative effects were reproduced.

Discussion

All of the most challenging reliable effects from Experiment 2 were replicated. Participants in both conditions spent reliably more time on their easier task. Giving-up times were regularly preceded by longer between-item times. Giving-up times were substantially and reliably higher in the harder of the two tasks.

The findings of this experiment offered clear support for the idea that duration of visits and giving-up times were influenced by task difficulty, but no support for the suggestion that people's decision to shift would be determined by the performance characteristics of the competing task. However, we must be cautious in the treatment of a null result: we are not claiming strong evidence against the influence of competing tasks.

Again, most participants allocated their time preferentially to the easier task, and presumably improved their overall performance by so doing. Of course this means that the total time spent on the Medium task was less when it was paired with an easier task than when it was paired with a harder task, yet the evidence from the within-visit timing data suggested that this difference resulted primarily from the different decisions to leave

Hard versus Easy tasks rather than different decisions in the Medium task caused by the characteristics of its competitor task.

Our two-process quantitative model, adding a subgoal-completion component to Green's rule so as to determine switch decisions according to number of items and time since last item, offered accurate fits to the participants' mean data, without special fitting of the three free parameters. Instead, parameter values were simply scaled from those that offered the best fit in Experiment 2. These scaled parameter values allowed good quantitative fits for the dependent variables in two independent groups of participants, each with different versions of the task. (Additionally, of course, the model explained all the important, replicated qualitative effects in the data.)

Although our model successfully explained longer giving-up times on more difficult tasks these may be considered counter-adaptive and may have resulted in too little additional time being allocated to the easier of the two tasks. However, they were only counter-adaptive to the extent that the tasks were not depleted. If, for example, there were no more words that could be generated in the easier task, then it would be rational to spend more time on the difficult task. None of the participants in these experiments came close to finding all the words from any of the sets of letters, but this is a different matter from their belief that they may have done just this. Perhaps, then, the higher giving-up times on the difficult task reflected the greater time required on difficult tasks to gain evidence that the patch had been depleted. Intuitively this is an appealing explanation of the phenomenon, quite distinct from the one we have been pursuing.

This explanation could readily be tested by providing participants with information concerning the extent to which they have exhausted the task. The next

experiment did exactly this, and at the same time extended our investigation to a similar but different task, in the name of generalization of the basic phenomena.

Experiment 4

This experiment used a task of searching for words in a grid of letters, a task that is quite popular in Puzzle magazines. As in earlier experiments, participants were free to switch between two puzzles, and were given the goal of maximizing the total number of words found. As a further incentive to attend to this overarching goal, participants were paid 10 pence for each word found.

A feature of the Scrabble task was that it produced responses that were sometimes clustered, both in terms of orthography and time (e.g. a burst of words with the same first syllable). We have not analyzed this clustering and consider it noise in relation to our hypotheses, which assume an idealized diminishing returns curve. We do not believe that the clustering could have caused the patterns in data to which we have attended, but it may have masked other patterns. The word search puzzle seemed likely, a priori to exhibit less clustering.

Another way in which the data from a word search puzzle would be purer is that it could not contain erroneous responses, such as misspelled or repeated words, which, for the Scrabble task, we have included in our analysis of giving-up heuristics but not overall performance.

Finally, an important assumption of our method is that we know on which task a participant is working when. We believe this assumption is quite strong in the case of the 7-letter scrabble tasks, but it is not inconceivable that participants occasionally thought of words for the non-active task. The word search task is even more strongly stimulus based.

The design of Experiment 4 was similar to Experiment 3, in that three tasks were used, with each participant receiving two of the three. However, the single task which all participants received was the easiest of the three. This allowed us a very direct test of a fundamental assumption of our account of Experiments 2 and 3, namely that participants were sensitive to the actual difficulty of each task, rather than merely the direction of relative difficulty of the pair. This assumption has already gathered support in terms of the durations of the very first visits, but the new design allowed a test in terms of overall time allocation. While we still predicted that both groups would spend more time in the easier task, we additionally hypothesized that participants in the Easy/Hard condition would spend less time in the alternative task than those in the Easy/Medium condition

The other novelty in Experiment 4 was a counter showing the number of words remaining in each puzzle. This ensured that participants had accurate information about the extent to which a patch (or puzzle) was depleted.

Method

Participants

Twenty four undergraduate students from Cardiff University participated in the study (5 male, 19 female; aged 18–22 years). Participants received a base rate of £2 for their participation. Additional pay was performance dependent (10 pence for every word found) and varied between £1.60 and £3.70 (mean of £2.48).

Design

Each participant received the same easier puzzle (Puzzle A: in which the hidden words were all names of Fruit and Vegetables) and one of two harder puzzles. The difficulty of the harder puzzle was either medium (Puzzle B: Birds) or high (Puzzle C:

Chemical Elements). Thus, there was one between-subjects factor (Group or puzzle combination) with two levels and one within-subjects factor (Puzzle Difficulty) with two levels. Puzzle order (or the identity of the initial puzzle) was counterbalanced within each puzzle combination group.

Materials

Three different word search puzzles were constructed: All puzzles comprised a 20 × 20 grid of letters, in which words from a particular category were present, some in each of four orientations × two directions (vertical, horizontal, both diagonals × forward, backward).

The difficulty of the puzzles was manipulated by varying the total number of words and the proportions of words in the different directions. Appendix 2 shows these puzzles and solutions. Word-frequency of target words was not explicitly controlled, but was likely to be higher in the Easy puzzle (fruit and vegetables) than in the Medium puzzle (birds) than in the Hard puzzle (chemical elements).

The software interface to the puzzle was programmed in MS Visual Basic 6.0. Within a 600 by 600 pixel puzzle grid participants were required to click on the first letter of any found word. If the clicked letter was indeed the first letter of a hidden word, a feedback sound was played and a status bar message congratulated them on finding a new item. The initial letter was then highlighted, and remained highlighted throughout the experiment. If the clicked letter was not the first letter of a word, an error sound was played and a warning message appeared.

A timer (counting down the number of seconds remaining in the experiment) was always visible in the upper left corner of the task window. Below this clock a labelled

counter showed the number of words left to be found in the current puzzle in red. Directly below this, the total number of words found so far was displayed in blue. In addition the number of error-clicks (false alarms) was displayed. Throughout the experiment, participants could switch between puzzles by clicking on a button in the upper right corner of the puzzle window (i.e., to the right of the counters). Each puzzle as well as the puzzle switch button was labelled with the semantic category of words it contained.

Procedure

On entry to the experimental laboratory, participants were presented with written instructions, and with a practice puzzle. Participants were explicitly instructed not to click on letters until they had found words, and told that more than a few such errors would lead to them being excluded from the experiment, and their fee being reclaimed. (In practice false alarms were rare, accounting for less than 5% of all clicks on the puzzle grid.)

The test phase lasted for 15 minutes. Upon pressing a start button either the Easy or the Medium or Hard puzzle was displayed (counterbalanced within groups) and participants were allowed to swap between puzzles whenever and as often as they chose (by clicking on the puzzle switch button). Thus, only a single puzzle was visible and active at any one time, but participants were free to switch between puzzles.

Results

Table 3 shows the main descriptive statistics of interest. Our main purpose in this experiment was to replicate the important significant effects of the previous two experiments, and the inferential tests will be presented with this in mind.

The number of task switches was somewhat lower than in the Scrabble experiments, but did not differ between groups in this experiment, $t(22) = .55$, $p = .59$, $\eta^2_p = .01$.

For number of words found, a mixed ANOVA (Group × Task Difficulty) yielded significant main effects of Group, $F(1, 22) = 7.6$, $MSE = 14.68$, $p < .001$, $\eta^2_p = .26$, and Task Difficulty, $F(1, 22) = 179.3$, $MSE = 13.20$, $p < .001$, $\eta^2_p = .89$, as well as a significant interaction, $F(1, 22) = 7.5$, $MSE = 13.20$, $p < .05$, $\eta^2_p = .26$. Both groups found about an equal number of words in the Easy puzzle, but, unsurprisingly the Easy/Hard group found fewer words in their alternative puzzle than did the Easy/Medium group, $t(15.5$, corrected due to unequal variances$) = 5.49$, $p < .001$, $\eta^2_p = .58$.

Across both groups, 21 out of 24 participants spent longer on their easier task (binomial test, $p < .001$). To test that time spent on a task depended on the absolute level of difficulty of that task we directly compared time spent on the Medium task versus the Hard task, between groups. This difference was reliable, $t(22) = 2.23$, $p < .05$, $\eta^2_p = .18$.

Each participant's very first task visit was analyzed to test for an effect of Task Difficulty independent of any knowledge of the competing task. Two t-tests were conducted (as each participant contributed only one value to either the Easy or harder task). For the Easy/Medium group there was no significant difference between very-first-visit times to Easy versus Medium tasks, $t(10) = .55$, $p = .59$, $\eta^2_p = .03$. For the Easy/Hard group significantly more time was spent on the first visit when it was to the Easy rather than the Hard task, $t(10) = 3.77$, $p < .01$, $\eta^2_p = .59$. The difference between conditions could, we suggest, again be understood in terms of the magnitude of the difficulty difference between puzzles.

The role of the competing task was again assessed in a similar way, by inspecting the mean duration of the first visit to the Easy task having seen the alternative task (each participant's second or third visit overall – one participant in the Easy/Hard group did not switch back to the Easy puzzle so provided no data for this analysis). T-test revealed no significant effect of competing task, $t(21) = -.45$, $p = .66$, $\eta^2_p = .01$.

Considering giving-up times (see Figure 7), data from four participants in the Easy/Hard group were excluded as they failed to produce any visits where they generated at least one item and then switched out of the Hard task. A 2 × 2 mixed ANOVA (Group × Task Difficulty) yielded only a significant main effect of Task Difficulty, $F(1, 18) = 4.51$, $MSE = 974.29$, $p < .05$, $\eta^2_p = .20$. There was no significant effect of Group ($F(1, 18) = .019$, $MSE = 951.29$, $p = .89$, $\eta^2_p = .001$) nor any interaction effect, $F(1, 18) = .28$, $MSE = 974.29$, $p = .60$, $\eta^2_p = .02$. Again, giving-up times were reliably longer for harder tasks. (Of the four excluded participants, one only switched once in the experiment, so did not generate any usable giving-up time data for the hard task. However, the other three participants were excluded because they failed to generate items on the hard task. These participants could be included if the total visit time for no-item visits is counted as a giving-up time. We made this adjustment for all participants and re-analyzed the data. Exactly the same pattern of significant effects emerged.)

Comparing longest between-item times with giving-up times (see Figure 7), a 2 × (2 × 2) mixed ANOVA [Group × (Timetype × Task Difficulty)] yielded a significant main effect of Timetype [$F(1, 18) = 6.03$, $MSE = 1473.51$, $p = .024$, $\eta^2_p = .25$] and a significant interaction of Timetype by Task Difficulty [$F(1, 18) = 15.30$, $MSE = 467.40$,

p = .001, η^2_p = .46]. As in previous experiments, between-item times were longer than giving-up times, particularly in the easier task.

The display of number of words remaining to be found suggested one final, novel analysis. Did participants typically switch into patches they knew to be richer, in terms of remaining items? If people were primarily motivated to switch into richer patches, the percentage of such switches should outweigh the percentage of switches into equal or sparser patches. As the instructions did not include any information about the actual richness of patches (overall or for any particular puzzle), participants had to have visited each puzzle once before making an informed comparison between their potential yields, so first visits were not considered. Also, since final visits were terminated by the experimental clock rather than a decision to switch into an alternative patch they were not considered in this analysis. Participants switched 52 times (53.6%) into a richer patch, and 45 (46.5%) into a poorer patch. The difference did not approach significance ($t(22)$ = 1.23, p = .23, η^2_p = .07. Switches into an equal patch did not occur.) Thus, it seems unlikely that the primary motivation for task switching was the prospect of a richer alternative patch.

Although the main purpose of Experiment 4 was to test the robustness of our experimental effects under different conditions and for a different task, we explored the success of our quantitative model at explaining the data. Because so few items were generated in the Hard task (M = 2.42) a high percentage of runs of the model failed to produce a giving-up time for the Hard task. Thus, we do not report fits for the performance of the group that received the Easy/Hard task.

The T value from Experiment 2 was scaled up to 48 seconds in accord with the longer task duration. (We modeled finding rates by a simple exponential, with initial probabilities of word-finding as .55 in the Easy task, .28 in the Medium task, each with a .976 decrement. The model was also extended to 150 time units each corresponding to 6 seconds.). Combining this with the other parameter values from Experiment 2 produced a reasonable fit to the data apart from the number of switches between the Easy/Medium tasks. However, given the completely different task we would not necessarily expect parameter values to be carry over unchanged. A best fit to the data was obtained using the values T = 60s, G = 30s, p = .1. The values of both these fits are shown in Appendix 1. For the best fit, the times on tasks, giving-up times and number of switches were all within the 95% confidence intervals of the participant means. Additionally, all the important qualitative effects were predicted.

Discussion

The most important findings from Experiment 4 were that the challenging complex of data relating to switch decisions was replicated with a different task, and despite participants being provided with veridical information about the extent to which each task has been depleted. In particular, participants generally chose to switch tasks rather frequently, and reliably devoted more time to the easier task despite reliably longer giving-up times in the harder task.

In addition to these replications, Experiment 4 yielded direct evidence that participants' time allocation was sensitive to the actual difficulty of each task, rather than based on a mere qualitative comparison. The decision to switch into a different patch was shown to be relatively unrelated to the remaining richness of the alternative patch.

General Discussion

We have reported three experiments in a new paradigm in which participants were free to allocate time preferentially across a pair of similar tasks, so as to try to optimize overall performance. Our interest has focused not on task performance per se, but on how the monitoring of task performance by participants informed their decisions to switch back and forth between tasks so as to manage their time effectively.

The most fundamental observation in this article is that when participants were free to allocate their time as they wished across a pair of tasks, most chose to switch between the tasks rather frequently. This bald observation raises two fundamental questions: why did participants switch back and forth between tasks, and how did they decide when to switch?

Experiment 2 suggested that participants switched partly so as to allocate time preferentially between tasks, according to monitored performance on the tasks. In all our experiments, the two available tasks had different gain functions, and these were not known in advance. In such situations, adaptive allocation must be done by switching, as it cannot be planned a priori. When the gain functions are dynamic, so that the most rewarding task at any moment may change, then relatively frequent switching is essential. These properties make the allocation problems we have studied distinct from typical choice situations, in which one task is always more rewarding than the others, or in which participants are required to select, rather than to allocate a continuous resource such as time. In these situations one might expect to observe task switching in order to sample (see Krebs, Kacelnik & Taylor, 1978). Future research will need to confirm the scope of

our findings across diverse tasks that differ in ways including the characteristic function relating performance gains to time on task.

A further aspect of our experimental tasks may have encouraged frequent switching. Because the time remaining to perform a task is likely to be inherently uncertain, even with a clock (which presumably can't be monitored continuously) it is rational to attempt to maximize total gain incrementally, in real time, rather than merely in total.

Despite the above discussion, it is interesting to note that participants who had a fixed separate budget of time for each task (Experiment 2) nevertheless chose to switch frequently, although less frequently than participants who were free to allocate their time preferentially. This suggests a general tendency to switch, independent of the attempt to allocate time adaptively according to monitored performance.

Following the optimal foraging literature we moved beyond statistics of overall time allocation to consider exactly when participants chose to move from one task to another to attempt some insight into how such switching decisions are made. Stephens and Krebs (1986) describe several rules-of-thumb that have been suggested to underlie foragers' patch-leaving decisions. The data from our experiments did not support any of these heuristics as stand-alone explanations for participants' switch decisions in the tasks that we studied. Problem solvers evidently did not simply perform each task for a certain period of time or until they had discovered a certain number of words (nor did they simply switch at random intervals, independently of their performance). If they had acted according to any of these heuristics, they would have failed to allocate more time to more rewarding tasks.

Participants were sensitive to the rates of reward within each task, but they did not use a simple giving-up time threshold. If they had used such a threshold, there would not exist between-item lags within a visit to a task that were longer than the giving-up time for that visit. Yet we reported that, on average longest between-item times were reliably longer than giving-up times (even if this effect did interact with task difficulty – it makes little sense to pursue a switching heuristic that could only apply in some tasks).

Finally, participants did not base their decisions entirely on the rate of return in the competing task, so as to switch to the currently more rewarding task (Experiment 3). Participants' switch decisions were sensitive to the current reward rate and to some extent independent of the competing task (for example, when the competing task had not yet been experienced, as indexed by reliable effects of task-difficulty on durations of very-first visits).

We have argued that the stochastic model that we have called Green's rule (following Stephens & Krebs, 1986) offers a good explanation of the qualitative effects in our data. This rule supposes that participants (foragers) set a time for which they are willing to remain in a task and increase this by a fixed amount with each success. This simple model applies readily to the ongoing interleaving of activity between a fixed number of tasks, even though it was invented to account for foraging behavior in which patches were encountered at random. The heuristic successfully allocates more time to the easier of the two tasks despite producing longer giving-up times in the harder task. In many ways it offers a simple and effective explanation of our data. However, it cannot fit the size of these effects – in particular the predicted difference between giving-up times

in the two tasks cannot approach the observed levels without failing to allocate sufficient time to the easier task.

Thus we extended Green's rule with an additional probabilistic driver of leave decisions. We noted that naturalistic studies of self-interruption report task switches immediately on subgoal completion (Gonzalez & Mark, 2005) - and we would argue that such a strategy is intuitively plausible. Further, and in keeping with this observation, the raw data of giving-up times in our experiments revealed a substantial minority of short giving-up times that Green's rule cannot produce. Thus we added a single free parameter – a probability of making a task switch on completion of a subgoal (i.e., after generating a word).

The resulting three-parameter model, with two independent bases for switch decisions is able to fit the size of the observed effects as well as their direction. Further, it actually predicts the effects, i.e. the model's behavior over the plausible range of its parameters is limited, and the participant data is within its scope, for a large number of dependent variables treated simultaneously. This success holds true in three experiments using two completely different tasks and the best fitting parameter values were consistent across the two experiments that used variants of the same task (Experiments 2 and 3). Thus we have not only presented reliable evidence for a challenging pattern of effects, but also reported support for our model that goes beyond most model fitting enterprises in experimental psychology (see Roberts & Pashler, 2000).

As far as we aware, the studies in this paper break new ground, and present some surprising findings concerning a rather universal behavioral tendency, i.e., discretionary interleaving of independent tasks. We hope that these studies have opened interesting

new avenues of research, developing recent attempts to extend foraging theory to human information processing tasks (e.g. Pirolli & Card, 1999). Discretionary time allocation among tasks is a topic of much interest in both theoretical and applied psychological communities, and this article suggests several possible routes of empirical and theoretical development.

References

Adamczyk, P.D., & Bailey B. P. (2004). If Not Now, When? The Effects of Interruption at Different Moments Within Task Execution. *Proceedings of the ACM Conference on Human Factors in Computing Systems of CHI'04*, 271-278. New York: ACM Press.

Allport, A., Styles, E.A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 421–452). Cambridge, MA: MIT Press.

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science, 15* 610-615.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory. Vol. 6*. New York: Academic Press.

Burgess, P. W., Alderman, N., Emslie, H., Evans, J. J., Wilson, B. A., & Shallice, T. (1996). The simplified six element test. In: B. A. Wilson, N. Alderman, P. W. Burgess, H. Emslie and J. J. Evans (Eds.) *Behavioural Assessment of the Dysexecutive Syndrome*. Bury St. Edmunds, UK: Thames Valley Test .

Charnov, E. L. (1976). Optimal foraging: the marginal value theorem. *Theoretical Population Biology, 9* 129-136.

Cnossen, F., Meijman, T., & Rothengatter, T. (2004). Adaptive strategy changes as a function of task demands: a study of car drivers. *Ergonomics, 47,* 218-236.

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

Gonzalez, V., & Mark, G. (2005). Managing currents of work: Multi-tasking among multiple collaborations. *Proceedings of the 8th European Conference of Computer-supported Cooperative Work.* Paris.

Green, R. F. (1984). Stopping rules for optimal foragers. *American Naturalist, 123* 30-43.

Herrnstein, R. J., & Heyman, G. M. (1979). Is matching compatible with reinforcement maximization on concurrent variable interval, variable ratio? *Journal of the Experimental Analysis of Behavior*, *31* 209-223.

Herrnstein, R. J., & Vaughan, W. J. (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits to action: The allocation of individual behavior* (pp. 143–176). New York: Academic Press.

Hockey G. R. J., Wastell D. G., & Sauer, J. (1998). Effects of sleep deprivation and user interface on complex performance: A multilevel analysis of compensatory control. *Human Factors, 40* 233-253.

Hodgetts, H.M. & Jones, D.M. (2006). Interruption of the Tower of London task: Support for a goal activation approach. *Journal of Experimental Psychology: General, 135,* 103 – 115.

Iwasa, Y., Higashi, M., & Yamamura, N. (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. *American Naturalist 117* 710-723.

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature 275* 27-31.

Maglio, P. P., Matlock, T., Raphaely, D., Chernicky, B., & Kirsh, D. (1999). Interactive skill in Scrabble. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, 326-330, Mahwah, NJ@ Lawrence Erlbaum.

Maylor, E. A., Chater, N., & Jones, G.V. (2001) Searching for two things at once: Evidence of exclusivity in semantic and autobiographical memory retrieval. *Memory & Cognition, 29,* 1185-1195.

McFarlane, D. C., & Latorella, K. A. (2002) The Scope and Importance of Human Interruption in HCI Design. *Human-Computer Interaction*, 17 (1), 1-61.

McNamara, J. M. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology, 21*, 269-288.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning. *Journal of Experimental Psychology: General, 131(3),* 349-363.

Miller, R.R., & Grace, R.C. (2003). Conditioning and learning. In *Experimental psychology* (A.F. Healy & R.W. Proctor, Eds.), Vol. 4 , pp. 357-397, of *Handbook of Psychology* (I.B. Weiner, Ed.). New York : John Wiley & Sons.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. NJ: Prentice-Hall.

Norman, D., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In Davidson, R., Schwartz, G., and Shapiro, D., (eds.) *Consciousness*

and Self Regulation: Advances in Research and Theory, Volume 4. Plenum, New York, NY. pp. 1-18.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review. 106*(4): 643-675.

Reader, W.R. & Payne, S.J. (in press). Allocating time across multiple texts: sampling and satisificing. *Human-Computer Interaction*.

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

Roberts, S., & Sternberg, S. (1993) The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience - A silver jubilee.* (pp. 611-653). Cambridge, MA: MIT Press.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124* 207-231.

Sandstrom, P.E. (1994). An optimal foraging approach to information seeking and use. *Library Quarterly, 64*, 414-449.

Sauer, J., Wastell, D. G., Hockey, G. R. J., & Earle, F. (2003). Performance in a complex multiple-task environment during a laboratory-based simulation of occasional night work *Human Factors, 45,* 657-669.

Shallice, T., & Burgess, P.W. (1991). Deficits in strategy application following frontal-lobe damage in man, *Brain, 114,* 727-41.

Smith, A.D., Glichrist, I.D., & Hood, B. M. (2005). Children's search behaviour in large-scale space: Developmental components of exploration. *Perception, 34*, 1221-1229.

Stephens, D.W., & Krebs, J.R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Styles, E.A. (1997). *The Psychology of Attention*. Hove, UK: Psychology Press.

Waage, J. K. (1979). Foraging for patchily distributed hosts by the parasitoid *Nemeritus canescens*. *Journal of Animal Ecology, 48*, 353-371.

FOOTNOTE 1. At the time these experiments were performed the UK/US conversion rate was £1 = $1.6 = 1.54 Euro

FOOTNOTE 2. Data from the Interleaving Free condition indicated 6 seconds was approximately the shortest plausible time within which participants both generated an item (typing and entering) and then switched to the other task.

Figure Captions

Figure 1. Experiment 1: Mean cumulative generation of words over time, for easy and hard letter sets (tasks), each fitted by an exponential (parameters given in text).

Figure 2. Combining and re-plotting the average data from Experiment 1 to compute the optimal allocation of time across tasks.

Figure 3. Experiment 2: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

Figure 4. Experiment 2. Range of difference between giving-up times (Hard task – Easy task) relative to proportion of time on Easy task when number of switches was fitted to participants' data. Participant data are also shown. The solid outline shows the 95% confidence interval around participants' mean. The dashed outline shows the 95% confidence interval around model data. Overlapping area indicates proportion of participants' data accounted for by Green's rule.

Figure 5. Experiment 2. Range of difference between giving-up times (Hard task – Easy task) relative to proportion of time on Easy task when number of switches fitted to participants' data. Participant data are also included. The solid outline shows the 95% confidence interval around participants' mean. The dashed outline shows the 95%

confidence interval around model data. The overlapping area indicates proportion of participants' data accounted for by combination of Green's rule and p-subgoal parameter.

Figure 6. Experiment 3: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

Figure 7. Experiment 4: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

Figure 1. Experiment 1: Mean cumulative generation of words over time, for Easy and Hard letter sets (tasks), each fitted by an exponential (parameters given in text).

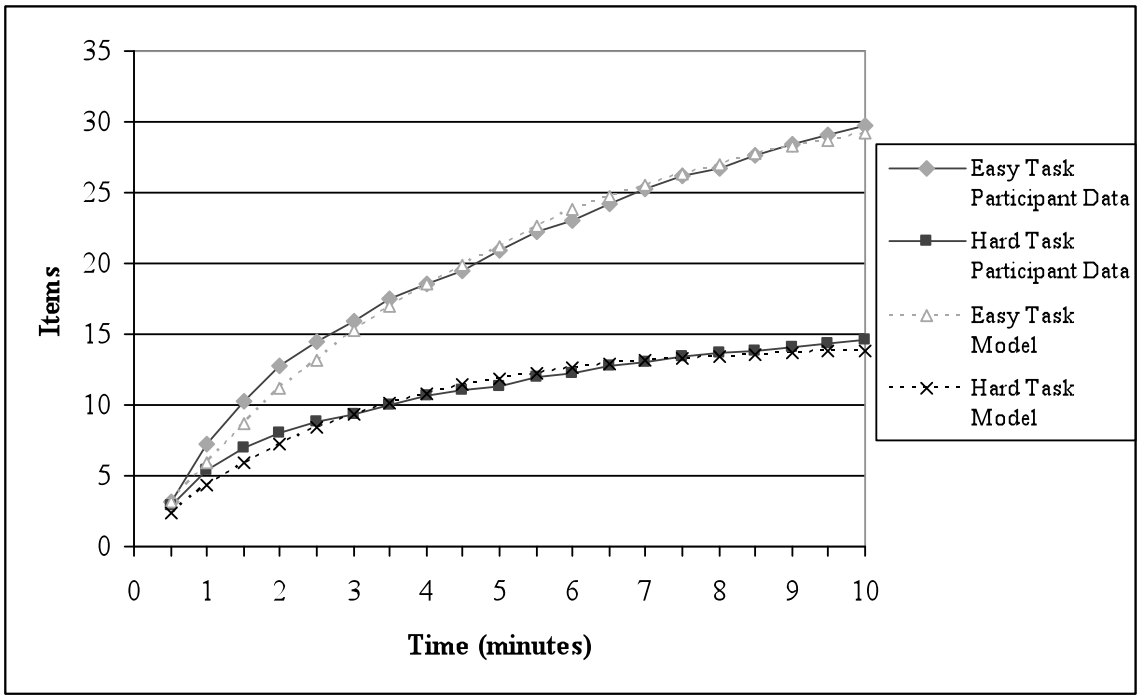

Figure 2. Combining and re-plotting the average data from Experiment 1 to compute the optimal allocation of time across tasks.

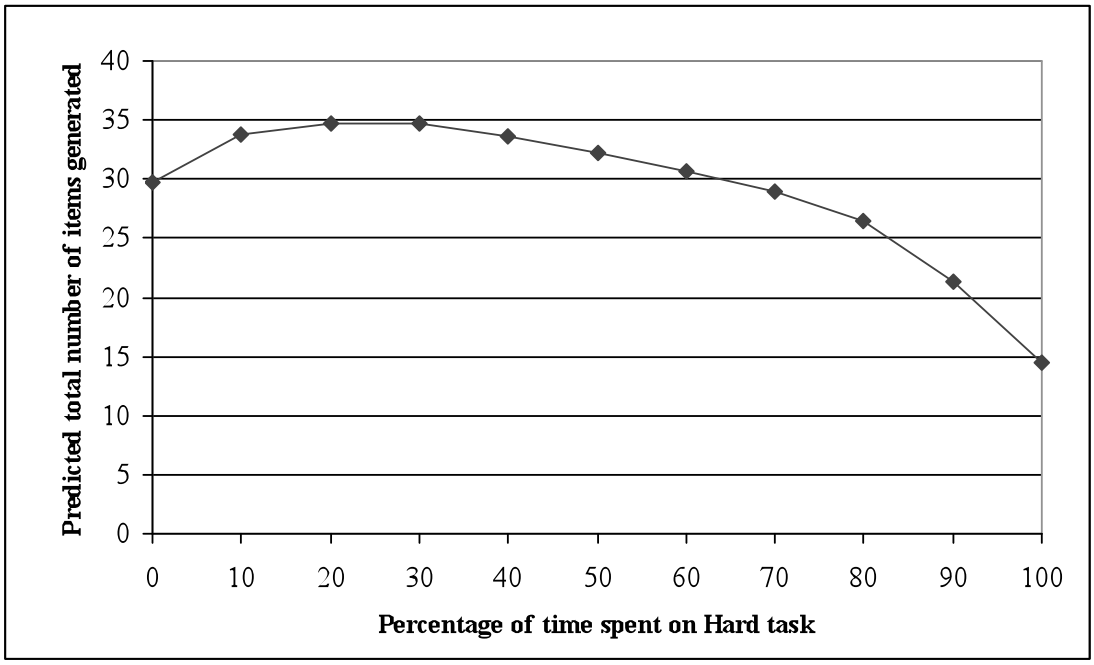

Figure 3. Experiment 2: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

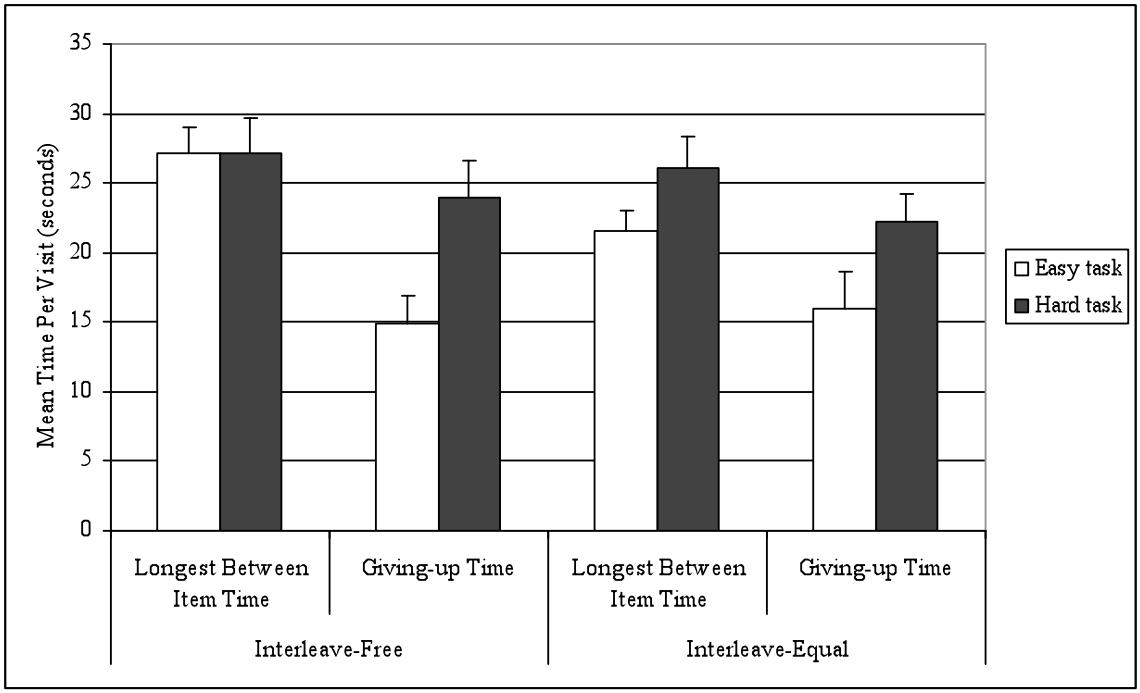

Figure 4. Experiment 2. Range of difference between giving-up times (Hard task – Easy task) relative to proportion of time on Easy task when number of switches was fitted to participants' data. Participant data are also shown. The solid outline shows the 95% confidence interval around participants' mean. The dashed outline shows the 95% confidence interval around model data. Overlapping area indicates proportion of participants' data accounted for by Green's rule.

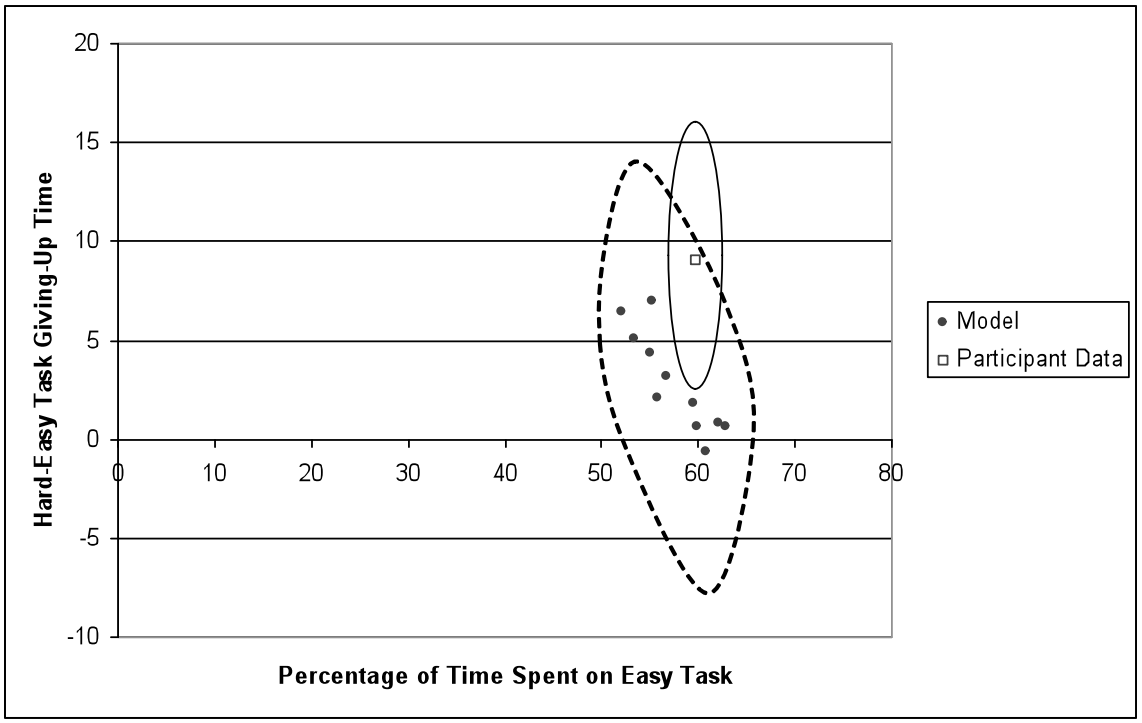

Figure 5. Experiment 2. Range of difference between giving-up times (Hard task – Easy task) relative to proportion of time on Easy task when number of switches fitted to participants' data. Participant data are also included. The solid outline shows the 95% confidence interval around participants' mean. The dashed outline shows the 95% confidence interval around model data. The overlapping area indicates proportion of participants' data accounted for by combination of Green's rule and p-subgoal parameter.

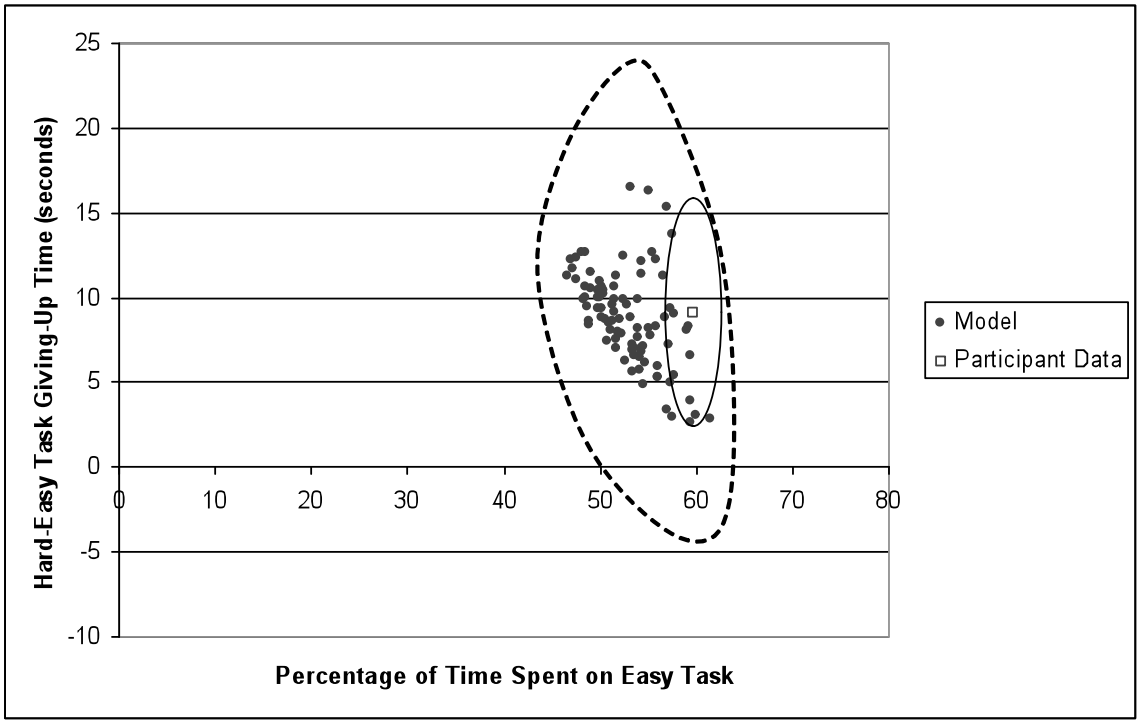

Figure 6. Experiment 3: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

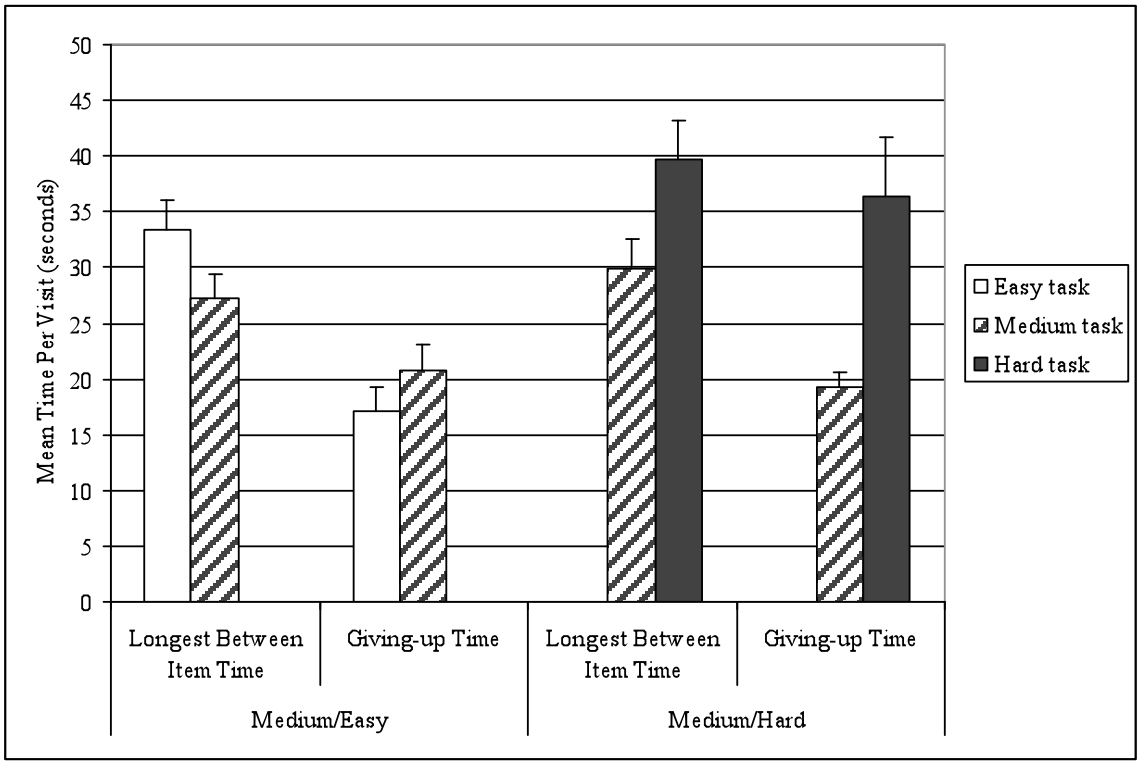

Figure 7. Experiment 4: Longest between-item times and giving-up times per condition. Error bars reflect standard errors.

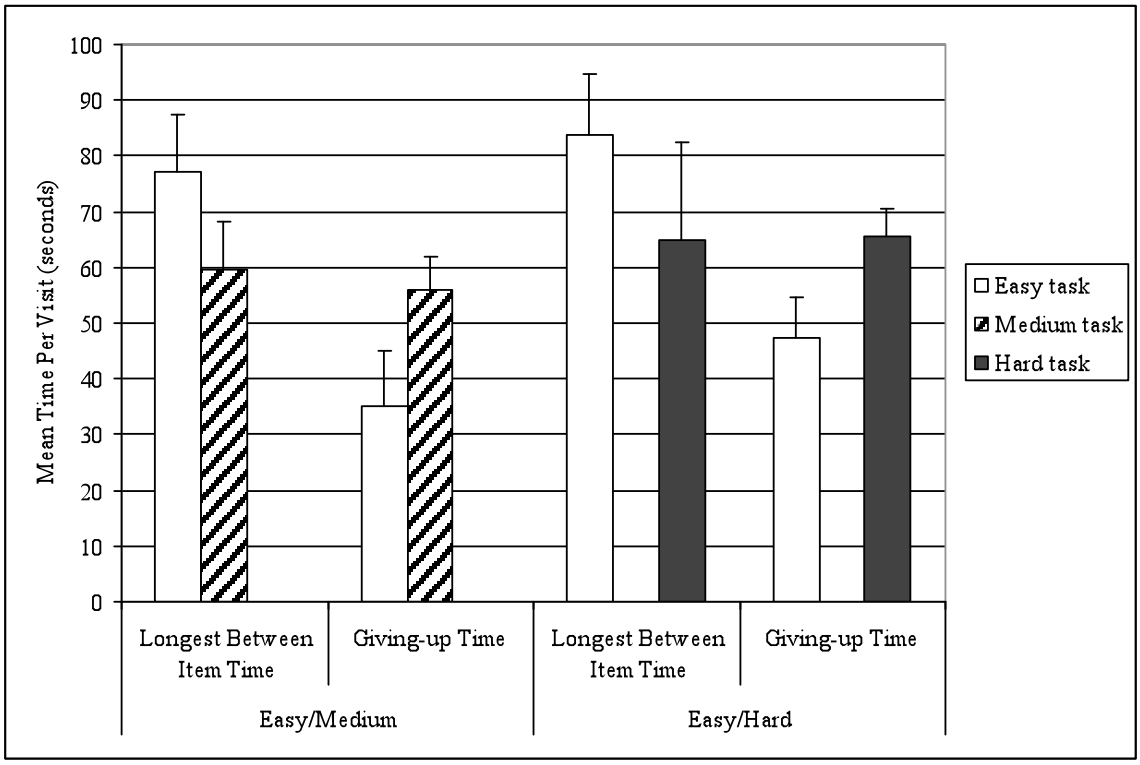

Table 1. *Experiment 2: Summary data.*

	Task	Interleave-Free		Interleave-Equal		Serial-Equal	
		M	SD	M	SD	M	SD
Words generated	Hard	7.46	2.50	10.04	3.04	9.63	3.28
	Easy	20.33	6.79	20.21	5.45	23.21	6.64
Time spent on	Hard	243.64	41.10	300.00	0.00	300.00	0.00
task†	Easy	356.36	41.10	300.00	0.00	300.00	0.00
Rate of return	Hard	1.88	0.69	2.01	0.61	1.93	0.66
(words/minute)	Easy	3.43	1.07	4.04	1.09	4.64	1.33
Number of switches		6.75	2.71	5.29	2.03	N/A	N/A

† For Interleave-Equal and Serial-Equal conditions time spent on task was enforced.

Table 2. *Experiment 3: Summary data.*

	Task	Medium/Easy		Task	Medium/Hard	
		M	SD		M	SD
Words generated	Medium	20.95	7.25	Medium	24.10	5.96
	Easy	26.10	8.71	Hard	11.75	2.84
Time spent on task	Medium	383.74	44.89	Medium	447.83	55.98
	Easy	456.26	44.89	Hard	392.17	55.98
Rate of return	Medium	3.28	1.12	Medium	3.24	0.72
(words/minute)	Easy	3.42	1.02	Hard	1.82	0.46
Mean duration of	Medium	112.88	71.36	Medium	125.08	76.22
visit	Easy	148.24	111.00	Hard	103.64	59.24
Mean duration of	Medium	128.85	99.38	Medium	156.66	101.83
very first visit	Easy	122.58	121.71	Hard	54.98	29.00
Mean duration of first visit to Medium task having seen alternative task		99.95	51.62		94.24	73.05
Number of switches		7.00	4.12		8.60	4.43

Table 3. *Experiment 4: Summary data.*

	Task	Easy/ Medium		Task	Easy/Hard	
		M	SD		M	SD
Number of switches		5.33	2.81		4.75	2.34
Words found	Easy	19.50	4.36	Easy	19.33	4.77
	Medium	8.33	3.39	Hard	2.42	1.56
Time spent on task	Easy	522.17	94.88	Easy	608.72	95.70
	Medium	378.42	94.87	Hard	291.87	95.72
Mean duration of visit	Easy	236.97	133.33	Easy	284.49	135.06
	Medium	161.58	82.02	Hard	104.87	50.74
Mean duration of very first visit	Easy	179.64	95.03	Easy	310.76	161.44
	Medium	143.66	127.32	Hard	54.25	42.22
Mean duration of first visit to easy task having seen alternative task		231.34	158.72		261.06	156.04