BINDING SERVICES
Tel +44 (0)29 2087 4949
Fax +44 (0)29 20371921
e-mail bindery@cardiff.ac.uk

Interfaces to Encourage Look-ahead: Impact on Problem Solving and Performance

Stephen R. Chambers

Thesis submitted to Cardiff University, School of Psychology, for the Degree of Doctor of Philosophy

February 2006

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ...Stephen Chambers..... (candidate)

Date ...01/08/06.....

Statement 1

The thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

Signed ...Stephen Chambers..... (candidate)

Date ...01/08/06.....

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan and for the title and summary to be made available to outside organisations.

Signed ...Stephen Chambers... (candidate)

Date ...01/08/06.....

Summary

The experiments reported in this thesis attempted to directly study the process of look-ahead during problem solving. Recent work has suggested that interface manipulations that increase look-ahead during problem solving lead to improvements in performance. However, evidence has been indirect, and there have been few attempts specifically made to quantify look-ahead span, changes that may occur over time and possible interactions with the task environment.

An initial experiment required users to specify 3 moves in advance while solving the 8-puzzle. The strict enforcing of look-ahead by even a small number of moves was unsuccessful in terms of improving problem performance. In fact, results indicated that such move enforcement may negatively affect performance.

Subsequent experiments, using both the 8-puzzle and Water Jars problems, provided participants with a motivation to plan using a Scoreboard system that rewarded greater planning and look-ahead. Results found this approach to be more viable, as the interface appeared to support the opportunistic planning behaviour frequently undertaken by participants. Across a series of experiments, increased look-ahead led to more efficient problem solving performance compared to controls, while leaving total time to solution unaffected.

Look-ahead span increased to approximately 11 steps when transforming the same start-state to a goal-state over trials on the 8-puzzle. When a new solution path had to be generated for each new problem start-state, look-ahead still increased over trials, but only to a span of approximately 4 steps. This look-ahead span was also observed during Water Jars performance when the Scoreboard manipulation was present.

A manipulation of 'system response time' (SRT) on Water Jars problems also led to improved performance but indicated an adaptation to the manipulation, leading to a lesser impact of SRT than previous manipulations. The results are discussed in relation to existing studies of planning, performance and the role that look-ahead may have in future studies of problem solving.

Acknowledgements

I would like to firstly thank my supervisor Professor Stephen Payne for his invaluable advice, encouragement and friendship over the last three years.

As always, I would like to say a huge thank you to my parents and family who have always been supportive of the choices I have made and for only ever wanting the best for me.

A thanks also to Guillaume, Laura, Kate, Peter and Alex who have always been supportive over the course of the PhD. A special thanks to Hans Neth, for the frequent programming assistance offered and given when needed most.

I would also like to thank Conor Saunter, Kelly, Stephen, Caren, Amanda, Colleen and Graham for providing me with support, a place to stay and cheering up when most needed.

Index of Tables

Index of Figures

Introduction

Experiment 1

Experiment 2

Experiment 3

Experiment 4

Experiment 5

Experiment 6

Experiment 7

Contents

Chapter 1

Introduction

"Human planning requires the cooperation of a number of cognitive processes including a look-ahead mechanism designed to generate multiple sequences of hypothetical events and their consequences, the development of stored structured event complexes that can guide movement from an initial to a goal state, execution linked anticipation of future events, and recognition of goal attainment."

Carlin, Bonerba, Phipps, Alexander, Shapiro & Grafman, 2000

As the above quote makes reference to, the process of 'looking-ahead' belongs to a much larger area of psychological study. Look-ahead is not only a key component of human planning but also one of the key processes underlying human problem solving. The ability to search one, two or more steps ahead in a problem representation that has been constructed and may only exist entirely in the mind is a high level cognitive skill.

Despite the importance of the role that look-ahead plays during human problem solving and planning, it is a process that has received very little direct study or theoretical attention. If it is implicated in problem solving behaviour it must therefore play a role in both performance and arguably to some extent learning. Increased knowledge and understanding about this important psychological mechanism may

have large implications for both problem solving research and also related applied fields such as Human-Computer Interaction.

More recent developments in problem solving research have seen a shift in the focus of study. Equal measure is now not only devoted to the study of internal cognitive mechanisms which typically operate during problem solving, but also to the effects that the external display and its interactions with internal mechanisms have upon observed performance (Payne, 1991; Zhang & Norman, 1994). This new dimension of study has opened up the possibility of measuring inherently internal mechanisms, like look-ahead, by manipulating the environment within which action takes place and recording the specific behavioural patterns that result. By asking more of a participant in terms of mental effort, the external display can act not only as a controller but a predictor of performance. The current research falls within this new domain of manipulating external displays to access information about important psychological processes involved in problem solving. More specifically, the current research focuses on two research topics in particular. If as recent studies have suggested but not directly tested, increased look-ahead leads to greater performance - can this be experimentally demonstrated? Secondly, if the first premise is correct, what are the typical features of look-ahead? Specifically, questions regarding average look-ahead span, increased use of look-ahead during problem solving and the effect of interactive mechanisms upon look-ahead span are all regarded as important research questions for the current work to provide answers to. However, if previous conclusions have been incorrect and the observed increase in performance was simply a result of increased care when selecting the next best move for example, then the look-ahead

element of problem solving may not have as great an impact on performance and learning as has been postulated.

The main body of the current introduction falls into three sections. Firstly, as look-ahead is a component process in an established field of study it will be beneficial to characterise the proposed role that the look-ahead process has within the process of human problem solving. Secondly, a discussion and analysis of the evidence for the typical span of look-ahead on a number of problem tasks commonly studied in the psychological literature will be given. Finally, from the varied body of research that has accumulated over several decades, evidence is examined that has found both benefits and null effects of increased planning during problem solving.

Problem Solving and Look-ahead

Conceptualising the look-ahead process within an established existing theory of problem solving will not only place it in a recognisable context but also allow the clarification of the typical functions that look-ahead performs during a problem solving episode.

The information-processing framework proposed by Newell & Simon (1972) described an analysis of problems in terms of their "problem space," that aims to provide an abstraction of the structure of any given problem. Any problem to be solved is often contained within a typical problem space from which there is a defined start state, with the aim being to travel through the problem space to the required goal state. The process of moving through the problem space is determined by the execution of a legal "mental operator(s)" or moves of which there can be any number

depending upon the problem, although the number of choices tend to be limited at least in the problems commonly studied in the literature. A subject may, through a look-ahead mechanism, mentally construct a new representation of the future problem state to aid in his/her choice of whether or not to continue with the implementation of the current 'best' choice or to postpone its execution and continue to generate alternative future states through the application of other competing operators. In theory, the depth of look-ahead can extend beyond the immediate successor state(s) to a state that lies any number of moves along the path through the problem space. Looking-ahead in a problem space to identify possible promising future states, identify possible future sources of difficulty and aid decision making when several seemingly advantageous states are available may have a direct impact upon performance or even the ability to successfully complete a problem task. The external display of the problem can in this time remain completely unchanged while the subject performs the required mental simulations. The choice and implementation of one particular operator over another will in turn lead to the arrival of a new state in the problem space. The exposure to new states, illegal states or revisiting past states all add to the often sparse pre-existing knowledge and representation of the problem. This continuous process of move selection and transformation from one state into new states further along the solution path generally continues until the end or goal state has been reached, although completion of a problem is obviously not always guaranteed. Through the construction of a graphical or abstract problem representation (see Figure 1 below), actual problem performance can be analyzed and compared against the abstract problem space. Such a comparison can allow further examination of possible problem solving strategies (Simon, 1975), reasons for possible sources of difficulty at particular points in the problem space (Kotovsky, Hayes & Simon, 1985; Kotovsky &

Simon, 1990) and the effects of practice on performance and learning (Gunzelmann & Anderson, 2003).

To help illustrate the size that a problem space can extend to even for a simple 3-disk Tower of Hanoi (ToH) problem see figure 1 below. The aim of the TOH is to transfer three separate discs of differing sizes placed on the left most peg of three possible pegs of identical heights to the right most peg with the following rules dictating which operators are legal:

1. A larger disk cannot be placed on top of a smaller disk
2. Disks can only be moved one at a time

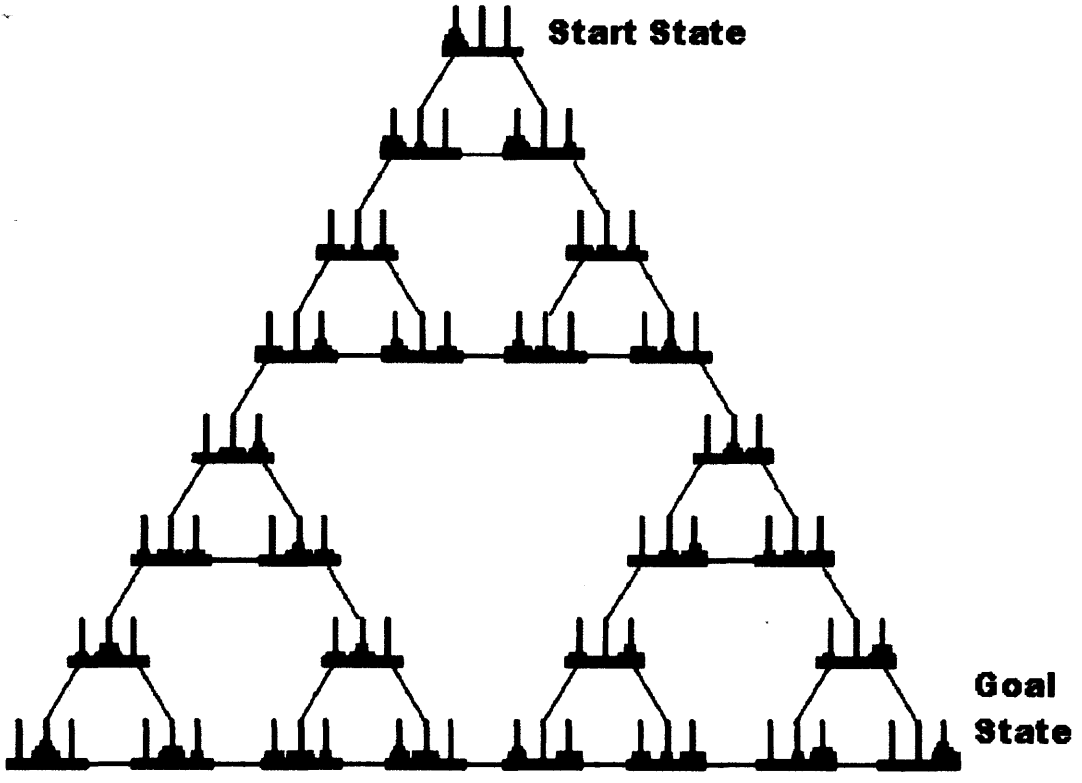


Figure 1. 3-Disk TOH Problem Space Representation

The 3-disc TOH problem can be solved in seven moves from start to finish. However, there are many alternative pathways of nodes that would still allow for completion of the problem but that would take many more moves. Identifying for example that the largest disk must be able to move unobstructed to the furthest peg would aid a problem solver to rule out moving the top disc to the middle peg as the next move would involve moving the middle disc to the furthest peg, thus blocking the target peg. With a look-ahead of 3 moves the intermediate state of having the smallest and medium discs on the middle peg, allowing for the largest disc to move unobstructed to the furthest peg would increase the likelihood of solving the problem optimally. A look-ahead component to problem solving should ensure, although not always, that new problem states are closer to the goal. Such a strategy would also enable the breaking down of problems into more manageable sub-problems which look-ahead may also operate within (Miller, Galanter & Pribram, 1960). Therefore, in light of overriding end goals the problem solving process may revert to smaller look-ahead steps while solving sub-problems that are more manageable.

It is important to point out that the size of a small novel problem's space can often be very large. Yet, even when the space is relatively small the solution to a problem can still be difficult to find and that problems with structurally identical problem spaces but different contents are not always equally difficult (Hayes & Simon, 1974; Kotovsky & Simon, 1990; Gick & Holyoak, 1980, 1983). Therefore, although the description so far has used the size of the problem space to describe the look-ahead process and implied a direct relationship between the size of a problem space, amount of look-ahead performed and problem solving success it is not a simple relationship

and the look-ahead process can only be considered one factor among many for observed problem performance.

Characterizing the Nature of Look-ahead

Like the proposed limited capacity of working memory (e.g. Atkinson & Shiffrin, 1968; Miller, 1956), evidence from the literature also suggest that human look-ahead is similarly limited, or at least in the problem solving tasks most commonly used in psychological study. Unlike productions models of performance that have a much larger working memory component (e.g. SOAR; Laird, Newell, & Rosenbloom, 1987) human processing does not organise information into indefinitely large "hierarchies of control". There are a myriad of reasons for only performing limited look-ahead and planning ranging from having insufficient knowledge or understanding of an often novel problem (Hayes & Simon, 1974), possible errors in the construction of the problem space (Lewis & Mayer, 1987), having incomplete awareness and memory of the environment's responses to proposed partial plans (Mayes, Draper, McGregor & Oatley, 1988; Payne, 1991), levels of experience in a domain (DeGroot, 1965), problem representation (Thevenot & Oakhill, 2005; Kintsch & Greeno, 1987), problem strategies afforded (Simon, 1975), individual preferences for greater planning and problem complexity (Davies, 2003), task goals or focus (Burns & Vollmeyer, 2002; Geddes & Stevenson, 1997; Vollmeyer, Burns & Holyoak, 1996), and simple natural limits on cognitive resources such as working memory (Cohen, 1996; Sweller, Chandler, Tierney & Cooper, 1990; Sweller, Mawer & Ward, 1983). The domains from which the limited nature of look-ahead is proposed differ substantially in both their levels of definition and classification, ranging from typically knowledge-lean problems such as the TOH to much more knowledge-rich

domains like computer and text editor use. Some of the studies discussed below make general characterizations and observations related to problem solving which have direct implications about the nature of look-ahead while others offer more quantitative data indicating the typical span, if any, of look-ahead.

From the space representation exemplified previously in figure 1, several assumptions may immediately be drawn about the role look-ahead may play. Firstly, from the size of the problem space created from such a simple problem it immediately suggests we are highly unlikely to have such a complete understanding or representation of the problem, thus affecting the ability to simulate states that we may not actually be able to construct due to incomplete knowledge of the problem. Hayes & Simon (1974) studied the effects that written instructions had upon subjects' initial understanding of a novel problem. Information was gradually assimilated and added to an expanding base of knowledge resulting in a gradually more detailed construction of the problem space. This process however takes time and inconsistencies between the actual problem space and the representation formed by the subject were often common. The 'understanding' process creates the problem space using a *language interpretation* component and a *construction* component. As the instructions are read and interpreted an initially sparse problem space is constructed. This usually contains basic information detailing the start and end states of the problem, key elements of the problem and the operators that can be applied in the problem space. The 'solving process' is implemented as soon as enough knowledge of the problem has been acquired to make the first tentative steps towards solution. Failures in the solving process lead to switching back to the understanding process where more knowledge is acquired that will once again aid the solving process. One such failure may be in cases

where the operators are incompatible with a problem state and the current set of operators are not applicable so as to bring about change in the current state. Not knowing what makes an operator legal in various situations has also been argued to be a major determinant of problem difficulty (Kotovsky & Simon, 1990). Therefore, look-ahead may be constrained by lack of knowledge or incomplete understanding but one implication is that look-ahead may increase with developments in understanding of the problem with time and experience. Therefore, problem solving behaviour may be reactive to increased experience with a task and it is plausible to assume that human planning and look-ahead will also change over the course of the problem solving task(s).

A second line of evidence that look-ahead proceeds in a limited fashion, rather than mentally searching through each of the possible alternative solution paths or nodes, comes from the frequent identification of weak or heuristic problem reduction methods or strategies commonly used (Pizlo & Li, 2005; Nilsson, 1971). These weak strategies aid in the management of cognitive resources by limiting the high demands placed upon working memory that computationally intensive planning or algorithms that guarantee success would entail (Anzai & Simon, 1979, Newell & Simon, 1972). Weak problem solving methods often form the basis for much of the problem solving performance typically observed during testing. Strategies such as hill-climbing, means-ends analysis, maximising gains and operator subgoaling, are all common strategies used to reduce computationally intensive cognition yet still allow for the completion of tasks with reasonable efficiency (Chronicle, MacGregor & Ormerod, 2004). With the current description of look-ahead contained within an information processing account comes several direct implications, prime among these being the

limited nature of human working memory (Baddeley & Hitch, 1974, Baddeley, 1986) and its likely effects on performance. It is therefore likely that intensive mental look-ahead cannot be performed without additional memory aids or the use of strategies used with the express aim of completing a task but reducing the demands on working memory.

Indeed, results from several TOH experiments (e.g. Simon, 1975, Anzai & Simon, 1979) have identified the *means-ends analysis* heuristic as one particular weak strategy typically employed by novice participants when solving the TOH problem. This strategy involves examining the distance from the current state to the goal state and selecting a particular operator that reduces the difference between the two states. If the intended use of the operator has not met certain preconditions due to rule constraints a small subgoal is identified to overcome the preconditions and implemented accordingly. The solution process can then continue through the selection and implementation of the operators judged to be the most likely to guarantee problem solving success. New subgoals may need to be identified again during the solution as and when needed and this general pattern of behaviour is thought to underlie the performance on many other problem solving tasks (Anderson, 1983). The reason such a method is used is because of the reduced reliance to hold many states in working-memory. A strategy that still allows near optimal solutions on problems while reducing working memory is also consistent with studies on bounded rationality (Simon, 1990; Goldstein & Gigerenzer, 1996).

A second example of using weak methods comes from the use of hill-climbing strategies, defined as the selection of one particular operator that always appears to

move the current state one state closer to the goal. This approach can be graded as 'steepest ascent hill-climbing' if an operator is chosen that always brings about the greatest advancement towards the goal. Such a strategy can be used but has also been shown to lead to problem solving performance that ensures that performance is at best prolonged and inefficient or at worst leads to a situation whereby a solution is never reached. An example comes from Ericsson's (1975) 'Single Tile Difference Model' strategy for 8-Puzzle performance. The 8-Puzzle requires the transformation of 8-tiles, contained within a 3 x 3 grid with one remaining space to allow moves, into a predetermined arrangement of tiles (Ericsson, 1974a, b, c). Subjects initially all used features of a limited hill-climbing strategy which always meant placing a tile in its goal position whenever possible. While this may work for one or two tiles at a time a hill-climbing strategy in this case will mean that getting three or more tiles in their correct sequence will likely never happen unless the problem is recognised before hand through planning. The more advanced strategies described by Ericsson (1975) such as the 'Row-Wise Subgoal Model' and the 'Distributed Attention Model' all lead to increased performance and contain greater degrees of planning and look-ahead.

A third line of evidence to characterize planning and look-ahead as both a variable and limited process comes from the classic study of planning described by Hayes-Roth & Hayes-Roth (1979). Using analyses of verbal protocols from subjects who were asked to plan their day in terms of completing a sequence of tasks to be completed, routes to be taken to complete tasks and the ordering of goals in terms of their importance, Hayes-Roth & Hayes-Roth instead regarded planning as a largely 'opportunistic' process. This approach differed from other planning approaches (e.g. NOAH; Sacerdoti, 1975) that typically formulated planning in terms of a number of

goals that consisted of a distinct hierarchy of increasing importance. A modification of these subgoals was implemented in an almost cyclic fashion until all goals had been altered into a definite and efficient plan. The opportunistic approach suggested by Hayes-Roth & Hayes-Roth (1979), argues that a plan can and often does start in what would appear to be a somewhat chaotic pattern. As new information is gained it is assimilated into the task feeding both higher and low levels goals (with lower levels affecting higher and vice versa) and with each new opportunity that might arise a new plan may develop or be modified in light of the newly acquired knowledge. Plans can be compared through a mental simulation to check for the efficiency of one plan over another or several and the results of these simulations will also feed into the ever changing plan. Plans develop in an incremental fashion, in "clusters", with a common knowledge 'blackboard' holding information, with plans seemingly adapting to changes in knowledge at different levels of abstraction. Although no attempts to quantify the possible depths that this planning extended to, this new conceptualisation observed that although look-ahead was limited, it still allowed for the generation of increasingly efficient plans that developed slowly over time.

Although increased planning and look-ahead may allow for better performance, there may be good reason why it is often only undertaken in a limited capacity. A distinction made by Rattermann, Spector, Grafman, Levin, & Harward (2001) differentiated between partial-order planning (similar to opportunistic planning) versus total-order planning. Partial plans are formed and revised as new information is acquired. This can be a very efficient form of planning and the description given by Hayes-Roth & Hayes-Roth indicates that it is the strategy most typically used by adults. The skill of assimilating new information into existing plans also appears to

develop with age. Younger subjects (7 – 11 years) appear to rely more on total-order plans that do not have the adaptability of partial-order plans (Ratterman et al, 2001). Flexibility allows the planner to avoid following a full plan from an early stage, instead enabling a subject to react to new or unexpected developments that may occur later in the planning process. The benefit of such a mechanism was demonstrated by showing that as the number of subgoals increased participants who typically used total-order plans had an exponential increase in their planning time with increased subgoals while partial-order planners had a simple linear increase in planning. One implication of such a result is that enforcing look-ahead to unnatural limits in the face of large problem complexity may result in either non-completion of a task or performance that is detrimental to what can be considered acceptable.

A number of recent problem solving studies have investigated the impact of task goals upon problem solving (e.g. Burns & Vollmeyer, 2002; Geddes & Stevenson, 1997; Vollmeyer, Burns & Holyoak, 1996), each of which are framed within a 'dual space' framework (Simon & Lea, 1974; Klahr & Dunbar, 1988). The results of these experiments have indirectly suggested that the amount of time spent searching a problem space by participants can be altered unintentionally by the goals that need to be accomplished. Moreover, it is also a process that can be conducted in the beginning with only the most minimal information about the possible problem states that can be entered into, yet still by the end of training and testing yield accurate solutions. Simon & Lea (1974) argued from a dual space perspective that participants are able to traverse either a 'rule space' by creating a representation of a possible state determined by the rules of the problem task. In the conceptualisation by Klahr & Dunbar (1988) this phase is known as the 'hypothesis space' whereby a new,

previously unexplored direction or space, is hypothesised to be reachable through use of the available knowledge or rules. A search of the secondary space that can then be explored is 'instance space', which allows the exploration of all problem states that can lead to the acquisition of new information for further exploration of the rule space.

The traversing of rule and instance space interacts, allowing the search and exploration of problem space which in turn leads to new knowledge and learning. Work that has manipulated task goals has given rise to the suggestion that increased search of rule space, manipulated by giving participants a non-specific goal (NSG) while using a system, leads to better understanding of how the system works and can lead to greater performance and transfer of knowledge to novel states. A specific goal (SG) will instead lead to greater search of specific instances, reducing or neglecting the search of rule space, which will in turn lead to lesser understanding and learning of the system. In a recent study, Burns & Vollmeyer (2002) used a 'Water-tank' control task that they describe as a 'linear' system. The mechanisms for operating the system involved manipulating values of the three different chemical inputs; Salt, Carbon and Lime to control three output measurements that were linked to water quality (Oxygenation, Cl Concentration and Temperature). Each of the inputs were differentially weighted to one *or more* of the outputs with the links or weightings not revealed to participants. The goal was to maintain the output measures of water quality by entering values from +100 to -100 on any of the inputs. Both sets of participants (NSG Vs. SG) had two rounds with which to explore the workings of the water-tank as well as a test phase whereby water output values had to be reached. A transfer phase was then also undertaken whereby output values that were inconsistent with previous values would also have to be attained.

Performance measures included a 'structure score' taken after each of the two exploration rounds measuring participants' understanding of the weightings between input and output variables. A 'solution error' score was calculated from the distance of their state to the required state and allowed calculation of the accuracy of their behaviour. A transfer phase then tested how accurate their system control was in maintaining the output levels on totally new output requirements and was calculated in much the same way as solution error. As expected, participants in the NSG group outperformed SG participants as indicated by higher structure scores, lower solution error measures and higher transfer scores.

A verbal protocol study was then conducted to directly test what NSG participants actually did while completing trials. Results from protocols found that SG and NSG groups during initial exploration phases partook in the same amount of hypothesis testing. The groups only differentiated during a second exploration phase where NSG participants increased their amount of hypothesis testing while SG participants dramatically reduced the amount of hypothesis testing and replaced it with goal testing. Therefore, prematurely exploring instance space instead of increasing or maintaining the amount of rule space searched, can lead to poorer learning of a task. Geddes & Stevenson (1997) even report evidence that a NSG manipulation can even lead to explicit learning in a task that has previously been studied as an implicit learning task (e.g. CLEGG; Berry & Broadbent, 1984; Broadbent, Fitzgerald & Broadbent, 1986) and to differentiate between implicit and explicit knowledge (Berry & Broadbent, 1987).

There are two main implications for the characterisation of look-ahead from the results of the above experiments. The conceptualisation of dual spaces is relevant for the current work as a similar dual process is also discussed in terms of 'planning'

versus 'action' while problem solving. Like the exploration of rule and instance space, planning and action may also interleave and spur problem solving performance forward. With greater time spent planning, increased numbers of possible new future states may be visited through mentally constructing the state, which may lead to greater performance through having explored and identified states of greater importance for performance and learning.

The second implication is that look-ahead maybe a process that can operate in a problem space, however inaccurately to begin with, and if given adequate time to flesh out the problem space, strong performance can still result. In the studies described above, performing look-ahead would initially be impossible as the relationship between changes in the future output states are completely unknown in response to any changes in the input values. Participants would have to track water levels on three different outputs with each move made, yet performance still quickly improved. With the gradual creation, exploration and refinement of the problem space, cued by the task instructions and increased time spent in the 'rule space', problem solving performance and learning can increase as a result. These results also suggest the universality of the look-ahead component for problem solving and its ability to play a role in almost any given problem situation.

Modelling the Look-ahead Component

While the previous examples have provided only characterizations and generalisations of the typical span of the look-ahead process, several studies have attempted to model and quantify the span of look-ahead on specific tasks. The first of these comes from the insight problem solving literature. Insight problems or 'ill-defined' problems (Ormerod, 2005), are a class of problem that appear to be quantitatively different from

typically well-structured problems that the current research aims to examine. There is evidence (e.g. Schar, 1996), that interaction or interface manipulations will not aid in the success of solving insight problems, so caution may be warranted in making direct comparisons of look-ahead span with such problems. In the 9-dot insight problem (e.g. Weisberg & Alba, 1981; Kershaw & Ohlsson, 2001), a series of dots arranged in a 3 x 3 grid must be joined by drawing four straight lines to connect the dots, without starting a new line from a place other than that where the last line finished. Performance is typically 0% on this task within a 5 minute time limit. Recent work on this problem by MacGregor, Ormerod & Chronicle (2001) estimated that participants' look-ahead values for 9-dot problems were best predicted by their model when set to 32%, 32%, 36% and 0% for look-ahead values of 1 – 4 steps respectively. These limited look-ahead values are one of the factors that they use to explain typical performance behaviour on this problem when solutions are being sought.

A second source of evidence whereby the look-ahead span has been quantified comes from a collection of studies that model the performance of subjects on 'Hobbits and Orcs' problems (Thomas, 1974; Greeno, 1974; Jefferies, Polson, Razran & Atwood, 1977), Water Jar problems (Atwood, Masson & Polson, 1980; Atwood & Polson, 1976) and TOH problems (Karat, 1982), although there is disagreement amongst these models as to both its span and importance for performance. The 'Hobbits and Orcs' problem involves transferring 3 members from each of their respective 'clans' to the opposite side of a river bank from where they are placed. There is a boat to aid the transfer that can carry a maximum of two members across the river at any given time. At least one person must be in the boat in order for it to cross the river. The final rule dictates that Orcs must never outnumber Hobbits on either river bank or the Orcs

will attack the Hobbits. Thomas (1974) and Greeno (1974) both incorporate a look-ahead function in their model each containing a small number of multi-step components of moves that account for typical performance on this task. The number of multi-step components is estimated to be in the range of three to four in line with the hypothesised number of key stages required during problem solving performance. Greeno (1974) found evidence that a group given 'corrections' during the acquisition phase, in terms of preventing backward moves and the provision of next moves performed worse after difficult states than groups which received little or no feedback and which therefore did not interrupt the look-ahead process at a key juncture. Performance patterns could not be accounted for by the number of external states in the problem but instead were predicted by a small number of internal cognitive changes involving two distinct phases in the problem. There were two to four stages in the first half of the problem and only a single stage in the second half of the problem. Further states were found to lack the Markov property, the assumption that the choice of one state will be independent of prior knowledge of that state. These small changes involved the development of plans or sets of moves with a small number of steps being contained within each change.

However, Jefferies et al (1977) adapting a model originally developed for Water Jars performance (described later), showed that performance could be modelled simply by incorporating a look-ahead component of only 1-step as opposed to the multi-step processes thought to be at work by Greeno (1974). The model consists of evaluation processes, memory processes for previous states and a 3-stage move selection process. Like their model for water jar performance the look-ahead is considered to have only a depth of one step. Only successor states from the current state are run

through an evaluation process and a move is selected in terms of its value from the evaluation stage, with moves being classified as either acceptable or unacceptable. If no move is chosen then a move not stored in LTM is searched for, i.e. not previously visited. If this new state is found and evaluated to be appropriate then it is implemented. If however no such move is found then the most optimal successor state is selected using information contained about the states in STM. This of course will be a limited number due to assumed processing constraints. In the event that no suitable optimal move is found then one is selected at random. Modelling performance on a number of river-crossing problems and their isomorphs, Jefferies et al. (1977) found that performance could be accurately predicted by their model. Rather than look-ahead processes being involved at difficult states it is simply failing of either memory capacity or being unable to find a move that meets criterion specified by its model.

Similar evidence of limited look-ahead comes from their process models of water jar performance. Atwood & Polson's (1976) model for Water Jars tasks similarly had no assumptions of any forward planning. The parameters underlying their model assumed severe 'data processing limits' and focused upon a simple evaluation of the next logical local operators rather than global features incorporating forward planning. Although taking some of the features of the general problem solver (GPS) in their use of a means-ends analysis heuristic they dismiss the use of setting up subgoals in a water jars task as a means to solving it, again mainly due to the limited processing capacity in short-term memory (STM). The use of subgoaling in water jars problems was also found to be inappropriate by Ernst & Newell (1969) as the problem cannot be easily decomposed into a series of subgoals, therefore limiting look-ahead

as a consequence. The ability to detect differences between the goal state and a current state are often difficult during water jar problem solving. They argue that memory limitations prevent human subjects from efficient forward planning. A more accurate description of human behaviour in this task can be obtained by applying means-ends analysis to local (rather than global) information and taking into account the subjects' limited memory capacity. In a further study Atwood, Masson & Polson (1980) implemented a 'move availability' condition which reduced working memory load by showing the next possible states available from the current state. Results showed no difference between their reduced memory condition and controls. In a second condition a 'Memory' condition which involved the move availability condition combined with an indication of which of those states had been previously visited. Although the new conditions led to significantly less moves the performance was still around five times the number of moves needed to solve the problems indicating that forward planning was still not being used to any significant degree and that small adjustments of their model could account for the new performance while still assuming a limited look-ahead mechanism.

Rather than attempting to discuss the merits of both approaches, the conclusions here are simply that if look-ahead is used then it is still a very limited component but that it may also be possible to solve a problem efficiently by simply comparing between the next best states and selecting what appears to be the most promising state. A constant look-ahead depth of only one step may therefore be typical of much problem solving behaviour.

Look-ahead and Expert Performance

So far the studies reviewed have involved studying novice participants in a novel domain. In the domain of chess skill the role that planning by looking-ahead plays in expert and novice performance has been extensively studied (e.g. deGroot, 1965; Charness, 1981, Gobet & Simon, 1996). Look-ahead is considered one of the key mechanisms allowing the identification of future states of the game, moves an opponent may choose and so on. It might therefore be expected that expert players (Grand Masters), with approximately ten years experience or more would have a much larger capacity for look-ahead. However, as important as the underlying mechanism may be Grand Master level chess players do not appear to look-ahead extensively. From the middle point of an average game there are typically 35 possible moves. Taking a look-ahead value of 3 legal continuous moves from this point onwards would result in 1.8 billion possible states and would increase dramatically if the search was at a deeper level (cf. Gobet & Simon, 1996). Therefore look-ahead processes, whilst needed during chess play, cannot account for the performance demonstrated by skilled chess players. Although differences do appear to exist in average depth searched there were no differences in maximal depth between grandmasters and less skilled players. Gobet & Simon (1996) argue that selective look-ahead to promising states is performed whereby a proposed template recognition process is invoked to identify known chess plays and that it is this process that is responsible for the performance typically demonstrated by experts. Such a strategy would also limit the load placed on working memory in accordance with the results demonstrated by novice participants described earlier. Similar results in terms of the amount of selective search have also been found by Charness (1981). Recent evidence suggests experts have not only a larger *visual* span but it is also the increased ability

to perceptually encode larger amounts of structured information that leads to greater performance, rather than increased memory capacity or a large look-ahead mechanism (Reingold, Charness, Pomplun & Stampe, 2001).

In information rich domains such as computer or text editor use Payne (1991) found that even heavy users of an everyday software package relied on information at hand constantly to aid in completing tasks. The studies showed that even experts or routine users of the software package did not have a full understanding of the behaviour of the system when completing higher level goals. Rather, the exact behaviour with a text-editor was only partially known and incomplete or simply wrong knowledge was quickly corrected by retrieving the necessary information from the display which would then immediately feed back into the task being completed. The lower level workings and interactions are not entirely known and therefore pre-existing entire plans are simply impossible to construct due to imperfect knowledge of how the system operates at all levels. This is even true of users that have years of experience in the environment within which they are working. They may have completed a task numerous times yet their knowledge is split between the display and general knowledge of how to complete the task.

By a similar vein, Robertson & Black (1986) found that users of a word processor begin by forming very short partial plans which when completed were followed by long pauses in action - often indicative of planning a new series of actions. The distribution of these pauses was also greatest between sets of 'superordinate' goals (Robertson & Black, 1986).

DISPLAY BASED PROBLEM SOLVING

The previous section described the often limited length that look-ahead may extend to while problem solving and the evidence clearly indicates that it does not seem likely that look-ahead can often be increased substantially due to logical or psychological constraints. The final two experiments described have also introduced the theme of performing actions using an external display as a key resource that can and does influence behaviour by the nature of the interaction (Davis & Wiedenbeck, 1998; Hutchins, Hollan & Norman, 1988), our reactions to it and the response(s) that we receive. While the conceptualisation has so far mainly focused upon the role of internal mechanisms, we should not ignore the fact that problem solving often takes part in an external environment. The problem environment can change dramatically during the course of problem solving and cognitive science has begun to study the importance of internal cognitive mechanisms combined with the importance of the external display which acts as the bridge between the mind and the results of the success of the problem solving process.

The importance of the external display was first brought to light by suggesting that depending upon the nature of the external display it could dramatically alter the success of the subject's performance. Larkin & Simon (1987) investigated the effects that diagrams can have upon problem solving success and argued that diagrams perform functions such as reducing the amount of search that is required when looking for particular features or information about the state of a task. A good diagram, one that does not conceal or misrepresent information critical to the task, can also aid in decision making by simply looking at the relationship rather than having to make more computationally expensive inferences which may be more prone

to error. Diagrams are however not an inevitable aid to problem solving as they can also be poor problem solving aids if their representation does not match the structure of the problem (Larkin, 1989). Larkin's (1989) Display Based Problem Solver (DiBS), is a production system with condition-action statements and a short-term memory component that is divided into two separate systems, one for internal memory and a second for items that are displayed in the external environment. Such a method negates the need to keep in mind large goal stacks and if applied to a simple ToH problem would be somewhat equivalent to using a perceptual strategy (for a detailed discussion, see Simon, 1975). Larkin argues that developing a model that accounts for the role played by external displays during problem solving can account for observations that previous models would be unable to account for due to the decreased reliance on a working memory component. Therefore, behaviour becomes more resilient against disruptions due to our offloading of task information to the external environment that is unaffected by interruptions or errors. Relying on the environment as an external resource may also ensure task performance remains relatively efficient as a completely new goal stack does not have to be recreated, although this may not necessarily be the case (for a discussion see Morgan, 2005). Performance can take its cue from the immediate state and proceed from there. This approach has mainly concentrated upon the external display as an information store and failures to display important features are also responsible for failures in problem solving.

More recently new approaches have included 'Distributed Cognition' (e.g. Zhang & Whang, 2005; Zhang & Norman, 1994). Rather than simply being an external memory source, external displays are much more fundamental to problem solving.

They are intrinsically entwined in the problem solving process. The problem space created from the internal representation created by the rules of the problem combined with the problem space created by the nature of the external representation does in fact create the abstract problem space within which we operate. The two spaces can be separated and analysed which in turn will allow a greater understanding of human problem solving. Our cognitive processes are distributed across these internal and external representations, the combination of which forms our overall concept of the problem at hand. The balance and makeup of both these spaces are the true reflection of our representation rather than simply our internal representation as previously reflected in early problem solving research rhetoric. Zhang & Norman (1994) showed that changing the number of internal/external rules for TOH problems and its isomorphs dramatically altered performance. Increased numbers of internal rules (i.e. when the number of internal rules constituting the problem space was high), resulted in performance that was more error prone and less efficient. When some of the internal rules could be distributed to the external environment, Zhang & Norman (1994) argued that it changes the problem space as well as reducing working memory. Increasing the number of rules to physically obvious structures increased performance and reduced the numbers of errors. Rather than being a simple memory aid however, they argue that information can be searched, information accessed and future states imagined to a much greater degree.

More recently Zhang (1997) has also included the role of affordances (Gibson, 1977) to the theory of external representations. Objects such as a particular problem representation for example, has properties that can immediately be accessed through perceptual processes and this can actually determine which rules are placed into each of the problem spaces that are proposed as forming the final abstract problem space.

Zhang (1997) argues that in some cases a 'representational determinism' (see also Cheng, 2002; Cheng, 1996) can result from certain objects properties and that this can also determine problem solving behaviour.

Norman (1988) argued that external objects can have a number of useful properties ranging from enforcing natural cultural constraints, reducing the degree of precision that is required from performance (see also Payne, 1991) and allowing increased monitoring of progress in the environment. Kirsh & Maglio (1994) for example have found that increased skill with the game of Tetris is often marked by experts rotating shapes to aid in the decision of where to exactly place an object. In fact, increased skill is often marked by an increase in the number of epistemic actions performed on the display (Maglio & Kirsh, 1996; Neth & Payne, 2002). Instead of performing all the necessary rotations simply through mental simulations, cognitive load is reduced by using the external environment as an additional problem solving aid. The ease at which rotations can also take place would presumably be a contributing factor in such behaviour. With the nature of external display based problem solving it appears then that performance should improve when changes that positively alter the amount of planning that is undertaken and particularly when the number of rules requiring internalisation is low.

Problem Solving Performance and Look-ahead

The following sections detailing the effect that increased planning and look-ahead have in relation to performance are categorised into two separate sections. This is due to evidence in the literature regarding the potential non-benefit to participants of increasing look-ahead. The second section details evidence from studies that

interpreted the results as indicating increased planning and look-ahead were responsible for the improvements observed in performance.

No Evidence of Improved Performance with Increased Look-ahead

Although look-ahead and planning processes are critical to most tasks there is evidence from studies using the Tower of London (TOL: Shallice, 1982) that increased planning results in no observable benefits in terms of problem solving performance. Originally developed to investigate the effects of frontal lobe damage to patients due to the proposed nature of the frontal lobes being primarily involved in planning (Shallice, 1988), the TOL has since been extended to test normal subjects by increasing the number of discs from three to five and equalising peg sizes to increase the number of moves and difficulty (Ward & Allport, 1997). Similar in appearance to the TOH, the TOL problem differs in several important respects. The discs which appear on the three pegs are actually of equivalent size, differing only in a superficial characteristic such as colour. Therefore, the restriction of disc size has been removed from the problem. A second key feature of the TOL as it is frequently implemented is that all moves *must* be pre-planned so that the solution path is known and subjects must be able to implement the minimum moves to solution in a rapid manner. Accuracy of solutions in the neuropsychological literature typically use the number of excess moves made as the measure of performance (Berg & Byrd, 2002). Given that look-ahead processes require the mental searching of moves in a problem space one would assume it to be an ideal task with which to study planning and look-ahead. However, and perhaps quite counter intuitively, the evidence appears to show little evidence of enhanced performance with increased pre-planning time.

Ward & Allport (1997) investigated performance of normal subjects on an adapted TOL task (TOL-R). Results showed that performance in terms of planning and number of errors made was significantly affected by the number of subgoal chunks contained in a problem. Ward & Allport (1997, p. 57) define a subgoal chunk as "a consecutive number of subgoal moves that transfer discs to and from the same peg". Furthermore, their analysis found differences of difficulty between problems containing the exact same number of moves, subgoal moves and subgoal chunks. They argued that in some cases move equivocation was absolute and a choice could be made. However, in other cases (when disassembling TOL problems) there were several competing goals vying for attention and resulted in longer planning and implementation times as a result (Carder, Handley & Prefect, 2004). Increased goal activations were to play a large part in problem difficulty. They argue against a simple working memory load explanation for problem difficulty (e.g. Baddeley, 1990). In fact Ward (1993) examined the performance of three different groups' performance on the TOL. One group planned the entire solutions in a typical TOL condition before implementing their solution. A second group were allowed to plan their moves by actively moving discs using the computer interface, thus reducing working memory load while planning. A third group were allowed to plan their solution on the interface and also did not have to implement their solution when they had reached a solution, again reducing working memory load while searching for a solution and while holding their plan during the implementation phase. Ward (1993) reported no differences between the second and third manipulations in planning time with increasing problem difficulty. There was an increase in time taken to implement the first move as the number of subgoal chunks increased, at a time where no other moves in a solution path had to be rehearsed or kept in memory. This was again given as

evidence that working memory demands may not be the critical factor for TOL-R problem solving performance (cf. Ward & Allport, 1997).

Phillips, Wynn, Gilhooly, Della Sala & Logie (1999) argue against Ward & Allport's proposal that working memory is not the restrictive factor for performance. Examining performance of TOL participants under dual-task conditions designed to load verbal, spatial and central executive processes (Baddeley, 1986), Phillips et al. (1999) found that dual tasks significantly decreased pre-planning times but did not alter mean inter-move latencies from controls. Numbers of excess moves made were more evident for tasks requiring the loading of spatial components and the central executive although these differences were not as significant as maybe would be expected. Articulatory suppression actually decreased number of excess moves made, implying that it prevented the use of an inefficient verbal rehearsal strategy. Pre-planning times for dual tasks were also unaffected by increasing problem complexity in terms of number of indirect moves, similar to a subgoal chunk but allowing more breadth of classification, and accuracy of solutions did not decrease with increased trial difficulty. No differences in execution times between participants in the dual-task group and controls also suggest that move choice(s) for dual-task participants was being supplemented in some way during the implementation phase. The argument is that for the TOL at least, planning is done on-line as moves are being executed. This characterisation of planning efficiently when the chance arises has been described before (Hayes-Roth & Hayes-Roth, 1979). The reduced performance when under executive and dual-task spatial loading tasks would seem to support such a conclusion. Phillips et al. (1999) suggest that the implementation phase of the TOL allows for any reduction in initial planning to be easily compensated for during the

move phase. One interpretation for the observed effects is that the interface or task itself is offering immediate cues which either suggest the next move immediately, indicated by problems with few moves, or offering a number of competing options that do increase time for normal subjects as they try to either verbally recite their initial plan or construct it anew from their current position.

Gilhooly, Phillips, Wynn, Logie & Della Sala (1999) investigated age differences in TOL performance. Participants with a mean age of 21.10 years planned entire solutions as well as older participants with a mean age of 66.95 years. Using the five disc TOL problem participants solved 20 different puzzles which differed in number of moves, number of unique solution paths to minimum solution and number of indirect moves (subgoal chunks).

Despite no difference in initial planning times, apart from the initial few problems and only a consistent difference in move implementation times, older participants did not differ significantly from young participants on the majority of the later problems in terms of number of excess moves to solution. Analyses of participants' verbal protocols made during the initial planning phase revealed that younger participants considered more moves on average, searched deeper in a solution path (6.0 moves versus 4.2 for older subjects) but this still did not lead to greater performance as indicated by number of moves on a trial by trial basis. Total moves made over the entire 20 trials did however in later analyses prove to be significantly different although it was not a large difference. A comparison between the accuracy of protocol plans with actual moves made also revealed significant differences in the accuracy of older participants' plans with young peoples' implemented plans matching much more closely their intended plans as extracted

from protocols. Protocol plans were also checked for the number of errors in plans for half of the trials that were more difficult overall. Number of errors increased with depth of search from depths of 2 – 3 levels and increased significantly from a depth of 4 moves onwards. This suggests that the depth of search for at least older participants is somewhere around 3 moves in the TOL task. Older and younger participants did not differ in the number of first moves considered indicating the search to have a very narrow focus (base moves, Gilhooly et al. 1999). Given the more error prone construct and reduced quality of their plans it still did not affect older participants' ability to complete the puzzles in the same number of moves as younger participants. Gilhooly et al. (1999) conducted a thorough examination of possible strategies that participants may use in the TOL. They argue that the main strategy used is a means-ends-analysis 'Goal Selection' strategy that requires participants to select a current move (as in a move selection strategy), and then currently choose an active goal that requires the removal of an obstructing disc. Although younger participants plans demonstrated significant differences in their depth, lack of errors and overall completeness the use of such a strategy by older participants ensures that when the move implementation phase begins they can use cues from the physical environment to concurrently activate goals and complete the puzzle effectively. Therefore testing of the move phase of the TOL-R may even be somewhat flawed due to the strategy that the problem seems to naturally afford at 'play' time to compensate for poor planning (Ward & Allport, 1997).

Yet the above evidence may not necessarily mean that pre-planning has no effect on TOL performance. More recent evidence has compared normal subjects' performance on the original TOL problem developed by Shallice (1982) with the Ward & Allport

(1997) adapted version which has been used by all the studies above. Unterrainer, Rahm, Kaller, Leonhart et al. (2004) actually found that longer pre-planning times were directly responsible for increased performance with greater planning resulting in more problems being solved error-free. Increased problem complexity still resulted in longer planning times as previously found. Contrary to previous TOL results however, short implementation times were evidenced by the high planning group suggesting that planners in the move phase of the original TOL were not relying upon the same spatial mechanisms responsible for performance previously found by Phillips et al. (1999). More efficient plans were implemented quicker than those who did not plan to the same extent. Unterrainer et al. (2004) argue that changing peg sizes to increase the number of possible moves to solution and increase difficulty which was the original motivation for Ward & Allport (1997) actually did more than that. It fundamentally changed the problem space that people work within when solving the TOL. When all pegs are equal sizes and a peg is not entirely full then actions can be performed on every peg. The numbers of alternative solution paths have been increased by making peg sizes equal. Using visuo-spatial, verbal and fluid intelligence tests, performance was examined and using a multiple regression methodology results indicated that only fluid intelligence as a significant predictor of problem performance. There was no effect of spatial mechanisms as previously found (Gilhooly et al., 2002). In a follow up Unterrainer, Rahm, Halsband, & Kaller (2005) compared TOL performance on an original Shallice (1982) version with unequal peg sizes with the adapted TOL problem. When all features were equal in terms of moves, subgoals, optimal solution paths, subgoaling patterns there were no differences in performance. However, in a second experiment when an original TOL was compared with a 5-disc version which had larger numbers of optimal solution paths as a

consequence of the physical changes, large planning and performance differences were found. Even though start states, goal states and number of moves and other structural features were identical the adapted 5-disc version took much less time to plan a solution and performance was significantly better than in the much more constrained original TOL task which had the added dimension of peg size. When all features of the TOL are identical there should be no problem with using the adapted TOL problem. However, with increasing number of moves the number of optimal solution paths appears to increase when peg sizes are equal which may account for the high levels of performance by non-planners in previous studies.

Further evidence of the complex nature of human planning comes from Davies (2003) involving the ToH task (Figure 1) and problem complexity. Investigating the number of trials required to reach a specified skill level of solving the problem twice in the minimum number of moves within a 120 second time limit. Davies (2003) found that with TOH problems with limited complexity (4-disc problems requiring 15 moves to complete), initial planners outperformed non-planners in total number of trials required to reach this performance benchmark. However, this effect disappeared with a 31 move 5-disc TOH problem with no effect of initial planning proving useful to complete the problem in the minimum number of moves. The amount of planning that participants may be able to do within a 15 second period and still complete a 31 move puzzle within the specified two minute period is not likely to be high. The design may have actually forced participants to simply start moving discs in order to complete trials according to the rules.

In a second experiment of planners and non-planners Davies (2003) examined individual preference for a particular strategy (High planners Vs. Low Planners). Having identified two groups of these high and low planners from a previous study they were given 4-disc and 5-disc TOH problems. The focus was to alleviate the possibility that subjects in the previous study forced into a planning condition would not have been able to cope due to it being an unnatural strategy for them. This more natural classification of subjects and simply allowing them to complete the TOH as they wished would still reveal differences but this time based on an indication of a participants' preferred strategy. There was once again an effect of complexity and a group by complexity interaction for total time with low planners completing the more complex 5-disc problem quicker than high planners. This effect was then reversed for the 4-disc problem with the high planners completing the problem faster than low planners. Similarly, with number of moves made, high planners only outperformed low planners at the moderate 4-disc level of complexity before this effect was reversed with the higher complexity 5-disc version. There was no effect of group on any of the analyses.

The studies described above all appear to show that planning does not necessarily lead to greater performance. However, as the current argument has tried to stress, human planning and acting do not necessarily operate in the fashion that the above studies have constrained subjects to operate under. However, there may be key differences between failures for planners who solve TOL problems versus those that attempt to solve TOH problems.

Firstly, both sets of problems can be solved online using perceptual strategies that can result in near optimal performance through problem reduction methods such as reducing problems into smaller subgoals and resolving those first before continuing. It seems undoubtedly true that the greater the number of subgoals (or subgoal chunks) contained in a problem the more difficult a problem appears to become (Ratterman et al., 2001; Davies, 2003; Ward & Allport. 1997). This does not necessarily mean that performance cannot be increased on these types of puzzles that contain such a structure. However, the success may be dependent upon the strategies that the problems themselves induce. From the TOL-R problems described there seems clear evidence that a spatial rehearsal mechanism is in operation during the implementation phase that can cover any shortcomings in pre-planning or incompletely specified plans. The longer on-line latency times reported in many of the papers for those that had reduced pre-plan exposure time suggest that this is a key issue for studies on TOL-R performance (Gilhooly et al., 2001; Phillips et al., 2001). It seems to suggest the need to develop a means of studying the quality of plans that is not just dependent on actual performance in terms of their number of moves above the minimum. TOL-R performance seems to be particularly confounded during the implementation phase where subjects can immediately draw on immediately available cues from current states and produce performance akin to extensive planners.

Failure of initial planning on the TOH problem once 5-discs were to be moved from one peg to another seems unsurprising given the short initial 15 second planning time allocated and the 31 moves required to solve the problem within the two minute time period. Such a stringent time limit may also have discouraged any attempts to plan during the implementation phase as the problem would simply have to be reset and

started again. It would be very unlikely that if even given a longer pre-plan time period that performance benefits would be witnessed on a 5-disc or greater TOH problem. Failure of planning in this example is simply related to complexity and possible initial planning benefits. Performance could be increased in this puzzle because of the different strategies that it induces that although can aid performance through perceptual cues do not invoke the same, seemingly powerful, spatial rehearsal mechanism that enables non-planners to enjoy reasonable performance on the TOL problem. However, in this case the manipulation would require planning to be allowed and more importantly encouraged *consistently* during the problem solving process.

One issue that directly leads from the above points is that problems which naturally induce or afford different problem solving strategies may also predetermine the success of any attempts to increase performance through interface manipulations. A problem that naturally affords an operator subgoaling strategy for example, will as a by-product, naturally induce look-ahead. Therefore, problems that do not afford look-ahead through either perceptually implicit cues about successor states or that instead naturally induce the weakest form of problem solving strategies (e.g. one-step hill-climbing) may be the problems that would benefit most from interface manipulations that aim to increase planning.

The current argument suggests that in line with more recent arguments of display based problem solving, increased performance may be best evoked by developing interfaces that support increased planning not only at an initial phase but also consistently while subjects are implementing moves concurrently or on-line. An

important feature of this may be that while aiming to increase planning and look-ahead, a certain amount of flexibility should still be allowed, thus permitting participants to decide adaptively when to increase their planning beyond a required minimum and when to increase planning and then act out plans as they see fit (Payne, Howes & Reader, 2001). These interfaces will also suit possible individual preferences for different amounts of initial and concurrent planning (Davies, 2003).

Benefits of Increased Look-ahead for Performance

Studies that have shown increased solving performance through greater planning have included the 8-puzzle (Ericsson, 1975; O'Hara & Payne, 1998, experiment 1 & 3, 1999 experiment 2), the slide-jump puzzle (O'Hara & Payne, 1999), water jars problems (Delaney, Knowles & Ericsson, 2004), river crossing problems (Knowles & Delaney, 2005) and Tower of Hanoi problems (Svendsen, 1991; O'Hara & Payne, 1998, experiment 4). The above studies with the exception of the Delaney et al. have increased the planning component in the plan-action continuum through manipulations that allow adaptive self-paced change during trials. The Delaney et al. study using water jars puzzles was the only study that asked participants to increase initial planning until a full solution had been reached. Vast differences between studies in their methodology, problem solving tasks and modes of interaction make direction comparisons somewhat difficult. The following section will try and give an account of the results of each of the studies with attention being paid in particular to the contribution that increased look-ahead may have had for observed performance.

Svendsen (1991) examined within the context of inducing different modes of learning (Berry & Broadbent, 1984) the effects that interaction style had upon problem solving

performance in a 5-disc TOH. Subjects specified and implemented move choices using either a mouse or keyboard command line driven interface. Similar to Davies (2003), subjects had to reach a specified level of performance of solving two trials in a row in the minimum number of moves. Unlike Davies' participants however, a two minute time limit per solution attempt was not imposed. Subjects stopped solving until two problems were consecutively finished in the minimum number of moves or until twenty trials had been attempted. Number of trials to reach criterion were significantly lower for those using a command based interface even on the complex 5-disc problem. Although Svendsen (1991) argued that the two interaction styles had induced different modes of learning (Hayes & Broadbent, 1988), it appears now more likely that the results were due to increased planfulness when considering moves. O'Hara & Payne (1998) argue that the command based interface would have shifted the plan/action balance to a more plan based approach and instead the results were better explained within a rational analysis framework. The keyboard command interface simply lead subjects to plan in greater amounts due to increased cost of implementing a move than to act in the rapid manner induced by a mouse driven interaction. Examining this possibility with all subjects using keyboard driven interaction but with cost of making a move varied the results confirmed their hypothesis. Subjects using a high-cost operator implementation interface took fewer moves across trials to complete a 5-disc TOH puzzle than those in the low-cost alternative (O'Hara & Payne, 1998, experiment 4). Trials to reach criterion were not used however in their manipulation as a performance criterion yet it seems reasonable to assume that the high-cost group would have reached such a performance criterion in fewer trials than those in the low-cost group. Rather than asking participants to plan over large periods or keep in memory large chunks of information the key difference

in these studies is that the interface induces planning in participants presumably in quantities that participants themselves find comfortable using. Increased planning is inherently induced by interaction style and allows a progressive adaptation to a more planful way of behaviour. This is achieved by increasing the quality of both the initial move choice and all other future moves while concurrently planning.

A recent study whereby participants were asked to plan the entire solution to a problem before being asked to implement their solution have shown benefits of increased planning. Using water jars problems Delaney et al. (2004) found that subjects performed significantly better than controls who had been asked to enter solutions using the minimum number of moves. Using a set of Water Jar problems taking from 3 to 7 moves to complete, Delaney et al. found that their plan group took less moves to complete puzzles than controls when asked to plan solutions in their entirety. The similarity of methodology between this study and previous TOL studies is comparable in terms of subjects being specifically told to plan the entire solution before entering their responses (e.g. Phillips et al, 1999). However and in contrast to those previous studies where no difference had been found between those entering moves concurrently and planners or between different age groups whose plans differed significantly in accuracy and depth (Gilhooly et al., 2001), there were significant performance differences. In line with the current argument one possible reason is that the water jars puzzles can not be solved by a simple online spatial rehearsal mechanism unlike TOL problems. No immediate cues as to the next best move are generally available from any given water jar state. The differences are also surprising given previous results from Atwood & Polson (1976) and Atwood, Masson & Polson (1980) who added a working memory load reduction and found no difference in performance or not as great a degree as would be expected. It has been

argued before that Water Jars performance cannot be broken into different numbers of subgoals like TOH problems (Ernst & Newell, 1969). It appears then that increased facilitation of performance for initial planning may be more likely to be observed when problems do not contain a subgoal structure, do not automatically lead to look-ahead but that would benefit from efforts to increase it. The evidence from recent cost of operator implementation studies are in favour of such an argument.

In a simple slide jump puzzle, again void of a natural subgoaling strategy, O'Hara & Payne (1999) used an operator lockout phase in which each time a move was made it caused a system response delay of approximately 4 seconds. This small delay was enough to induce greater performance from those using the delay interface than those whose responses were immediately implemented and allowed to enter their next move without delay. Similar to the 8-puzzle results the argument put forward was that the cost of this delay increased planfulness in the lockout time group.

O'Hara & Payne (1998) using an 8-puzzle originally studied in detail by Ericsson (1975) also argued for increased look-ahead and planning through an operator implementation cost manipulation. As previously discussed, Ericsson (1975) identified that a common hill-climbing strategy typically characterized novice performance. The problem involves tiles arranged in a 3 x 3 grid which must be arranged from a start state into a specified goal arrangement. Again, using operator implementation cost as a proposed means of increasing planning participants once again in the high-cost operator group outperformed those in a low implementation cost group. This effect transferred to users who then went on to use a low cost interface. They had maintained their performance suggesting that they had learned how to solve the puzzle differently than those in the control condition. A similar

implementation cost effect using the 8-puzzle has also been found by Golightly (1996). A verbal protocol study taken from participants using either the high or low cost interfaces were coded and showed that the high cost group made significantly more 'plan' statements than those in the low-cost group. It still remains to be seen if such an interface increased look-ahead per se, or simply increased the consideration of the next best moves. Again, the 8-puzzle does not contain a natural subgoal structure and cannot be simply broken down by mean-ends analysis heuristics.

It appears that characteristics of the problem task and its environment, i.e. how we interact with a particular problem, can affect the degree and the success that planning and look-ahead will have. Largely, problem choice may play a large role in determining the success that any interface manipulations will have. From the evidence of studies that have found performance benefits a number of predictions can be made:

1. Look-ahead manipulations will work best with problems that do not contain mechanisms that allow near optimal performance through control or online performance.
2. Manipulations to increase look-ahead must contain a certain degree of flexibility to mirror the seemingly adaptive nature of planning
3. Forcing look-ahead will not always work if the task is overly complex
4. Increasing look-ahead should also lead to greater learning benefits

This leads to the first experiment in the current thesis which adopts the 8-puzzle as used by O'Hara & Payne (1998).

Chapter 2

Look-ahead Manipulations and 8-Puzzle Performance

Introduction

The first experiment is an initial exploratory test of recent explanations that specific interface manipulations can lead to increased problem solving performance as a result of inducing greater planfulness and look-ahead (O'Hara & Payne, 1998). This more planful approach to problem solving was demonstrated by more efficient performance in terms of solving problems in fewer moves and by the end of testing, requiring less time to complete trials. If increased levels of planning and look-ahead were the mechanisms responsible for better performance then directly increasing the amount of look-ahead that is required by participants, in order to bring about change in the external display, should increase performance and provide some qualification to the proposed mechanisms stipulated within the rational analysis framework by O'Hara & Payne (1998).

Experiment 1

As discussed in the introduction, it may be that increased performance will be more likely when applying interface manipulations to puzzles that do not afford spatial rehearsal from the display Given these observations and recent success with 8-puzzle manipulations it was decided to initially begin testing the effect of look-ahead manipulations with the 8-puzzle.

The 8-puzzle is an ideal puzzle for comparing manipulations as it allows for a large number of possible start states, goal states, problems that differ vastly in their solution length and to investigate possible transfer effects (cf. Ericsson, 1974a).

A second advantage is the large number of performance measures that can be recorded from 8-puzzle performance such as total time to complete trials, total moves to solution and inter-move latency times. In addition to these measures O'Hara & Payne (1998) identified "palindromic" move sequences as a useful measure of solution efficiency. A palindromic move sequence in the 8-puzzle is indicative of undoing previously made moves as it immediately returns the problem to its previous state. If for example the move sequence '**78338**632' were performed in the 8-puzzle its structure contains a palindrome consisting of 4 items which means that only 4 moves (those not in bold) were actually responsible for the new state arrived at as opposed to the 8 moves made in total. Evidence of larger numbers of palindromes contained within a solution path may indicate poorer levels of planning (Delaney & Knowles, 2005).

To further clarify how palindromic move sequences were classified and counted in the current experiment, consider the tile sequence '1**75335***78448*721'. There are circumstances when a second palindrome may lie within the span of an already existing palindrome as the above example demonstrates. The sequence shows how one palindrome can overlap into a previously identified palindromic sequence. When such sequences occur, the current classification determines that a tile can only be counted once within a unique palindromic sequence. Therefore, the above example contains two palindromes, one with a length of 6 items (in bold) and the second containing 4 items (in italics). The second occurrence of tile 7 is counted within the

first palindrome and so is excluded from being counted again in any future palindromic sequence that immediately follows.

The enforcing of a number of moves to be specified and implemented per 'move' required the setting of the minimum number of moves look-ahead users would be required to specify before changes in the physical problem state could occur. Given the apparent limited natural look-ahead of participants indicated by previous studies the limit was set to asking participants to move only 3 moves or more per entry. Increasing the number of moves to be considered at any one time may have detrimental effects upon performance by increasing the load on working memory, although the current limit in itself may also be high. Given the large number of possible states available in the 8-puzzle when considering even a small number of steps ahead this caution seems warranted. Previous modelling work (e.g. MacGregor et al., 2001) also suggests that 4-moves are unlikely to be considered by participants and may indicate a natural limit for naïve participants on a problem solving task. The allowing of entry lengths greater than 3-moves would also allow greater performance *if* participants choose to or are able to look further ahead.

Furthermore, the look-ahead condition in the current experiment allows for the measurement of numbers of moves entered at one time, above and beyond the enforced minimum. It would be expected that with increased task experience and practice in a domain, increased levels of planning and performance across 8-puzzle trials should be observed (Gunzelmann & Anderson, 2003).

Based on the previous success of interface manipulations a number of predictions can be made regarding the effects that plan based manipulations may have on performance. The requirement to plan a number of steps in advance should force participants to consider greater numbers of options and may lead to the discovery of optimal solution paths over trials – therefore a reduction in the number of moves needed to solve the puzzles over trials. It would also be expected that inter-move latency times will be much greater to begin with for look-ahead interface users as this is typically observed in situations where greater planning is taking place (Robertson & Black, 1987).

If the current manipulation works as intended there may also be the observation that as experience with the interface and perhaps quality of plans increase, that there will be a reduction in latency times for both groups. Whether look-ahead interface users can reach levels of performance comparable to 1-Move interface users by the end of trials remains unclear. A similar pattern for total time to complete the puzzle should also be evident. More efficient solutions as they are generated will compensate for the increased time spent planning and reduce total time to solution by the end of trials.

If previous cost-related interface manipulations are correct then less undoing of moves just made should also be a consequence of the increased planning. The ratio of palindrome sequences should be lower in a planning interface although again this prediction is not certain. A high cost of undoing a move may reduce the proportion of backtracking moves but in an interface where there is no associated cost, as in the current manipulation, perhaps this behaviour will not manifest itself. Subjects using a look-ahead interface may be actually more likely to undo previously made moves as

they off-load the task demands to the external display encouraging greater exploration and resetting of moves.

Method

Subjects

Forty-Four Cardiff University undergraduates ranging in age from 19 – 34 (Mean age = 22.05 years, S.D. = 2.37), took part in the experiment for either course credit or a payment of £5. All participants were given full money or credit upon finishing the six trials encouraging sustained performance across trials. Six participants, two from the 1-Move condition and four from the Look-ahead condition, were excluded from the experiment as they were unable to complete the required number of trials and requested to leave the experiment.

Design

The experiment involved two between subject factors which comprised of Interface which had two levels and problem start state which also had two levels (A or B). The interface manipulation involved a 1-Move control group that specified and implemented a single tile move per press of the 'Return' key. The look-ahead interface condition required participants to specify three or more sequentially legal moves that could logically follow one another in the problem space. For both interface groups a move(s) were only implemented if the move or sequence was legal in its entirety.

The second between subject factor of start state, being either problem state A or B (see Figure 2a & 2b below), each taking 17 moves to transform into the goal

state (see Figure 3), were randomly assigned to participants. Equal numbers of participants in each condition received either state.

The within subject factor of trial had six levels, that were completed one after the other. Dependent measures of total number of moves to solution, total time to solution and inter-move latency times were all recorded by the computer program on each trial. The ratio of palindromes to total moves and number of moves minus palindromic sequences were calculated from participants' performance data when the experiment was completed as well as number of moves entered per return key press for the look-ahead interface group.

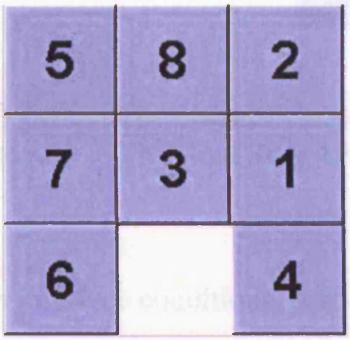
Materials

The 8-puzzle was created using Visual Basic 6.0 Professional. The interface and method of control were identical for all participants. The only change in terms of user interaction with the puzzle was the number of tiles that participants could specify and move at once.

For both Interface conditions a command line with which participants used to input their moves was positioned directly underneath the 8-puzzle. Those in the 1-move condition were prevented from entering any more than one move by the computer program. Participants could input and delete a move into the command line as many times as they wished. A move was only recorded upon the pressing of the 'Return' key and its subsequent validation.

The look-ahead condition required the specification of a sequence of 3 or more logically sequential moves that could be implemented. The program prevented them

from entering fewer than 3 moves, unless moves to final solution required fewer than 3, or from changing the current problem state until the entire sequence of moves had been specified by the participant and each move in the sequence validated by the program.



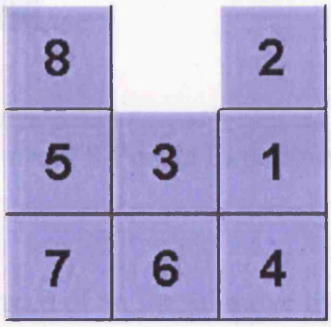


Figure 2a. Start State A (17 Moves) *Figure 2b.* Start State B (17 Moves)

For example, if a participant was attempting Start State A (Figure 2A above), a valid sequence of moves would be to enter '412' and then press the 'Return' key. The moves are implemented in the order specified by participants. The overriding goal of all participants was to transform their given start state into the goal state (Figure 3 below), a picture of which was also positioned on screen and which remained constant throughout all trials.

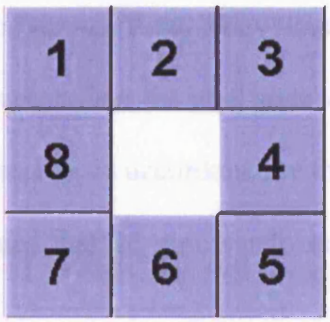


Figure 3. The Goal State Used in All of the Current 8-Puzzle Experiments

For both interface conditions, participants were informed of an illegal move by a message box and told to check their previous move and enter a valid move. Look-ahead interface users would have to make sure that their 3 or more move sequence was correct. If an illegal move occurred at any point in their sequence they were prompted about an illegal move and the problem state was returned to the arrangement prior to move entry.

Procedure

Participants were seated in front of a computer and informed that they would be taking part in a problem solving experiment. The 8-puzzle was shown to participants on the screen with a novel start state, not used in the current experiment, and informed that their aim was to transform the start state into the goal state which was shown on screen. Furthermore, they would be required to solve the 8-puzzle over six trials. Participants were informed that the only keys they would need to use would be the numeric keys 1 to 8 which corresponded to the 8 tiles of the puzzle and the 'Return' key. The 1-Move participants were shown the controls and told to complete the puzzle in as few moves as possible. Participants in the look-ahead condition were

instructed to enter 3 or more moves at a time. They were also informed that they could not enter less than 3 moves, unless the goal state was less than 3-moves away, and that all moves had to be possible in accordance with the order that they were entered. They were also informed that the aim was to solve the puzzle in as few moves as possible.

All participants were given a simple rotation practice problem involving moving 6 tiles into position to make sure they understood the task and controls. Once this had been completed and participants confirmed they clearly understood the task they could start the experiment by pressing a button marked 'Begin'. When a level was successfully completed participants began the next trial by once again pressing 'Begin' until all 6 trials had been completed.

Results

The dependent measures were log transformed to stabilise for variance. Data were analysed using a 3-way mixed ANOVA. There were no effect of problem start states (A or B) on any of the problem solving measures and so will not be discussed further.

Total Moves

The effects of interface on total number of moves over the 6 trials can be seen in Figure 4 below. Analysis on total number of moves to solution revealed, contrary to predictions, no main effect of interface on performance, $F(1, 36) < 1$, *ns*, MSE = 07. The analysis did reveal a small but significant effect of trial on total moves, $F(5, 180) = 2.719$, $p < .05$, MSE = .20. The results were modified by a significant trial x interface interaction, $F(5, 180) = 2.83$, $p < .05$, MSE = .21.

Simple main effects analysis revealed that the 1-Move control condition solved the 8-puzzle in fewer moves than the look-ahead condition at trial 1, $F(1, 36) = 7.12$, $p < .05$, MSE = .08. However, at trial three the initial increase in performance by 1-Move controls had reversed, with performance of Look-ahead interface users performing more efficiently than 1-Move interface users, $F(1,36) = 4.35$, $p < .05$, MSE = .08. The performance on the remaining trials revealed no other performance differences.

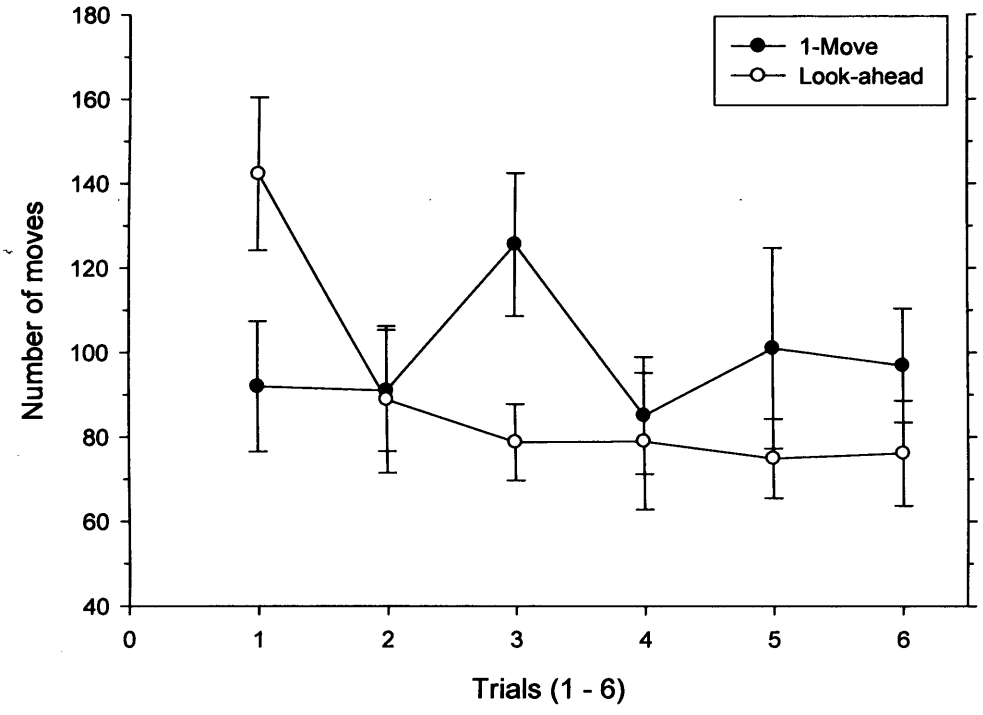


Figure 4. The Effects of Interface on Total Moves to Solution

Trend analysis over trials revealed significant linear, $F(1, 18) = 14.98$, $p < .001$, MSE = .91, and quadratic components, $F(1, 18) = 4.82$, $p < .05$, MSE = .43, for those in the Look-ahead condition. Trend analysis on total moves over trials for those in the 1-

Move group however, revealed neither linear, $F(1, 18) < 1$, *ns*, MSE = .01, or quadratic curve components, $F(1, 18) < 1$, *ns*, MSE = .04.

Total Moves minus Palindromic Sequences

Figure 5 below shows the total number of moves minus palindromic sequences.

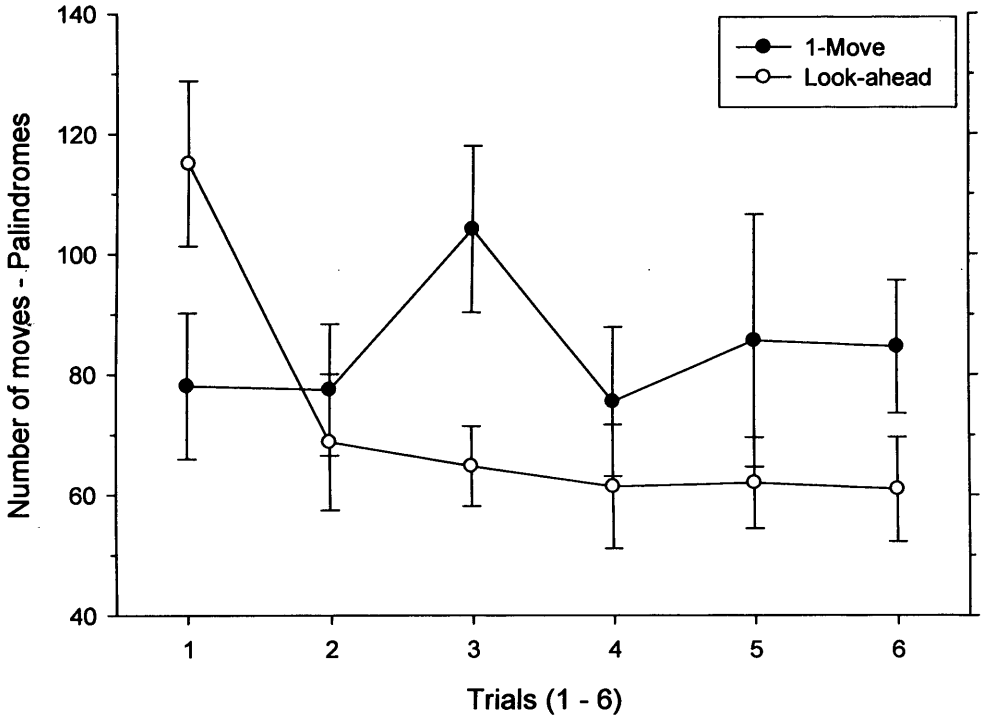


Figure 5. Total Moves Minus Palindromic Sequences for both Interface Conditions

As before, there was a significant effect of trial, $F(5, 180) = 2.97$, $p < .02$, MSE = .19, and a significant interaction between trial x interface, $F(5, 180) = 2.99$, $p < .02$, MSE = .19. As reported for number of moves there was no significant effect of interface on number of moves minus palindromes ($F < 1$), suggesting the number of palindromes was not significantly different between groups.

Simple main effects analysis revealed a significant effect at trial 1, $F(1, 36) = 6.48$, $p < .02$, MSE = .07, and a reverse effect of interface at trial 3, $F(1, 36) = 4.59$, $p < .05$, MSE = .07. No other effects of interface on number of moves made were significant at trials 2, 4 and 5 (All F's < 1) although trial 6 indicated a slight trend for less moves in the look-ahead condition but this was not significant, $F(1, 36) = 2.04$, $p < .16$, MSE = .09.

A Trend analysis on performance over trials for 1-Move interface users revealed no linear or quadratic components to the curve (F's < 1). Trend analysis on Look-ahead performance revealed significant linear, $F(1, 19) = 17.73$, $p < .001$, MSE = .83, and quadratic trends, $F(1, 19) = 6.00$, $p < .03$, MSE = .40, to the curve.

Ratio of Palindromes

The ratio of palindromic sequences to total number of moves was calculated and is presented below in Figure 6.

Contrary to previous findings there was no between subject effect of Interface on the ratio of palindromic move sequences to number of moves made, $F(1, 36) = 2.31$, $p > .1$, MSE = .06.

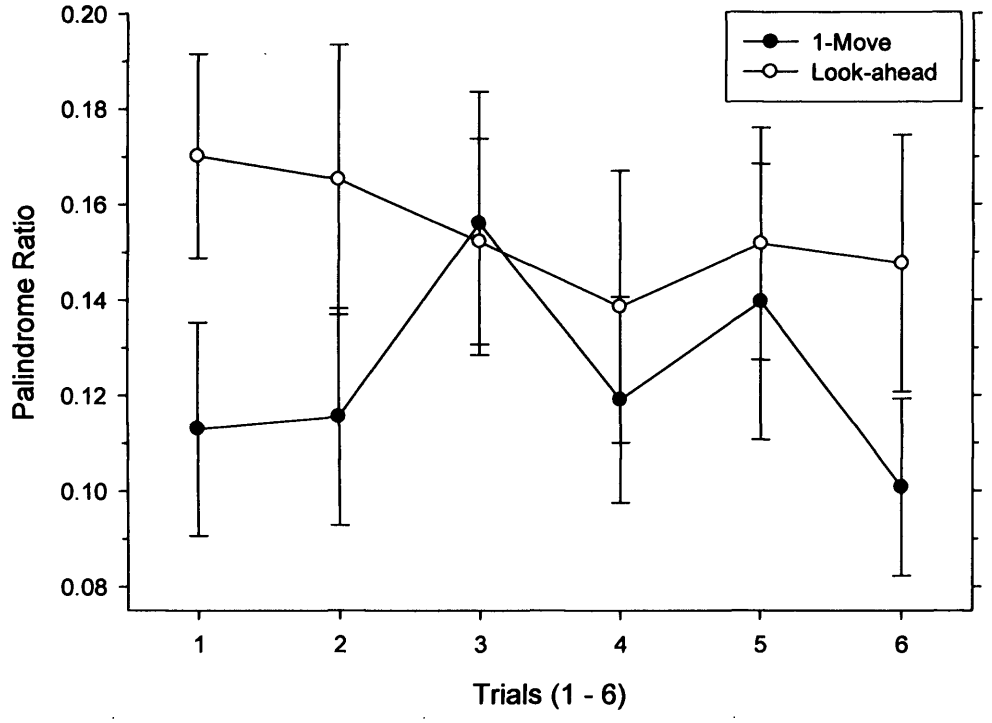


Figure 6. Ratio of Palindromes to Total Number of Moves for both Interface Conditions

There actually appears to be a trend for the look-ahead interface users to have more palindromic moves than controls although this difference was not significant. There was no within subject effect of trial or interactions with interface (F's < 1).

Trend analysis revealed no linear or quadratic curves for the look-ahead group (All F's < 1). There was also no linear (F < 1) or quadratic curves, $F(1, 19) = 1.97$, $p > .1$, MSE = .02, for 1-Move group performance.

Total Time

The effect of interface on total solution times over trials can be found below in Figure 7. The main analysis revealed that interface had no effect on the total time taken by participants to complete the 8-puzzle, $F(1, 36) = 1.56$, $p > .1$, $MSE = .33$. The analysis revealed a significant effect of trial, $F(5, 180) = 7.61$, $p < .001$, $MSE = .66$, with both interface groups completing the 8-puzzle in less time by the end of the 6th trial. The ANOVA also revealed no significant trial x interface interactions, $F(5, 180) = 1.49$, $p > .1$, $MSE = .13$.

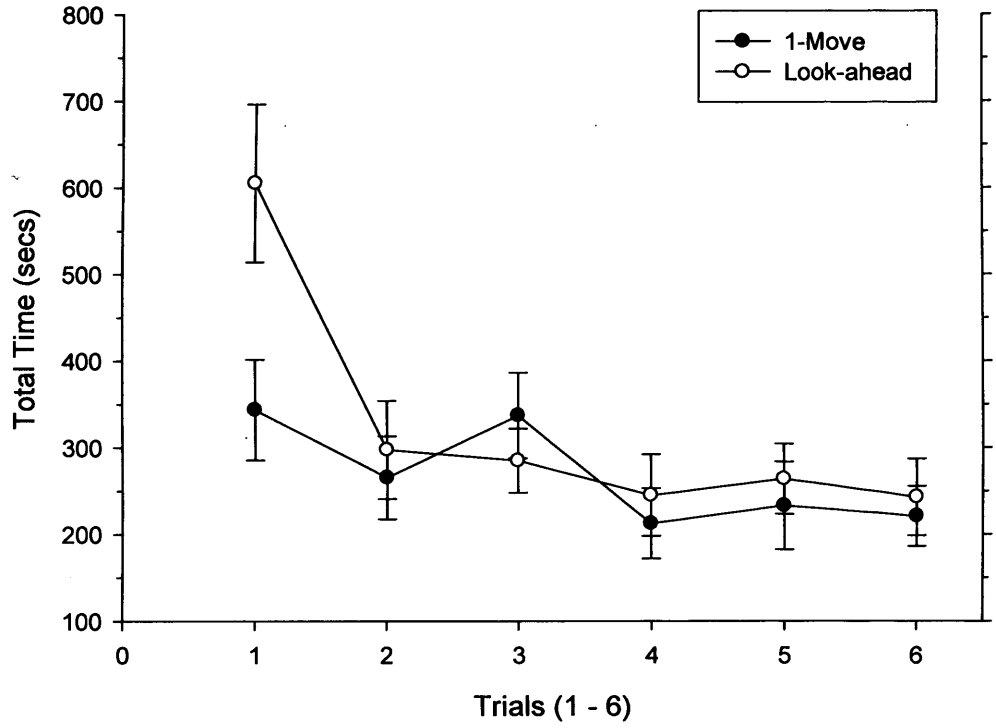


Figure 7. The Effects of Interface on Total Time to Solution Across Trials

Trend analysis revealed significant linear, $F(1, 18) = 19.57$, $p < .001$, $MSE = 1.79$, and quadratic components, $F(1, 18) = 4.84$, $p < .05$, $MSE = .54$, for Look-ahead interface users. For the 1-Move condition there was also a significant linear component, $F(1, 18) = 4.73$, $p < .05$, $MSE = .75$, but no quadratic curve component ($F < 1$).

Inter-move Latency

The inter-move latency times for both groups can be seen below in Figure 8. As predicted, the analysis latencies for both groups revealed a significant main effect of interface with Look-ahead interface users taking significantly more time per move than 1-Move users, $F(1, 36) = 7.75$, $p < .01$, $MSE = .71$.

There was also a significant effect of trial, $F(5,180) = 23.42$, $p < .001$, $MSE = .16$, which was moderated by a significant trial x interface interaction, $F(5, 180) = 2.75$, $p < .05$, $MSE = .02$. Simple main effects analysis revealed no differences between the two interface conditions at trial 1 ($F < 1$), or trial 2, although this did approach significance, $F(1,36) = 3.878$, $p < .057$, $MSE = .07$. Significant differences were found however between trial 3, $F(1, 36) = 7.16$, $p < .02$, $MSE = .16$, trial 4, $F(1,36) = 10.33$, $p < .01$, $MSE = .15$, trial 5, $F(1, 36) = 10.74$, $p < .01$, $MSE = .25$, and trial 6, $F(1, 36) = 7.56$, $p < .01$, $MSE = .16$, latency times.

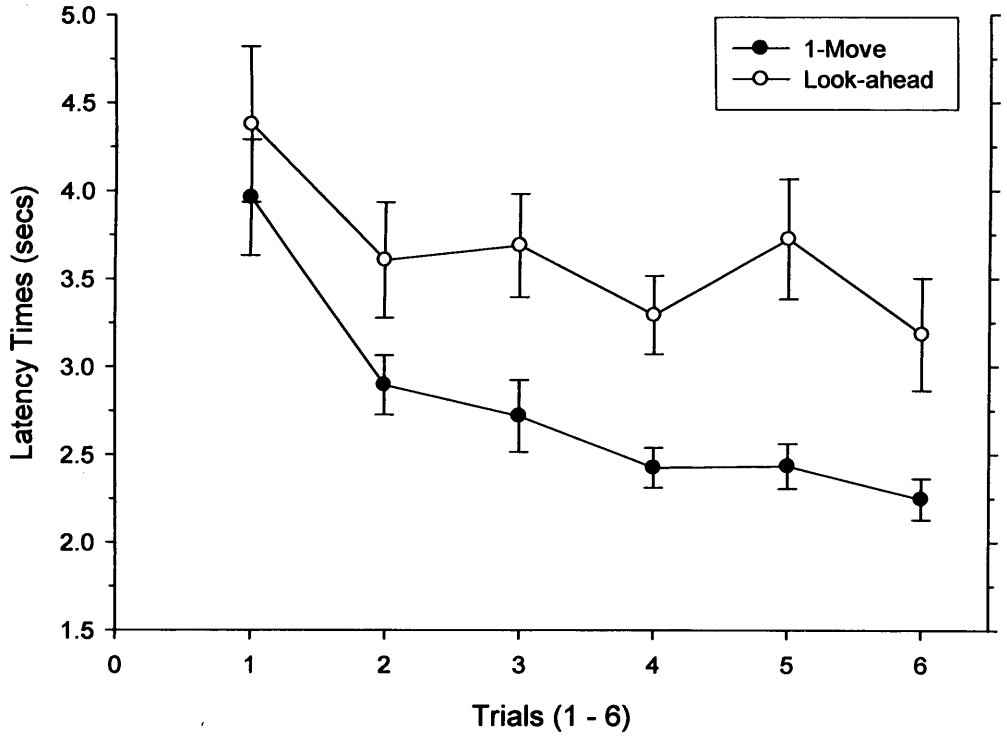


Figure 8. The Effects of Interface on Inter-Move Latency Times

Trend analysis on latency times across the 6 trials for the look-ahead group revealed a significant linear component, $F(1, 18) = 9.41$, $p < .01$, $MSE = .15$, but no quadratic components to the curve ($F < 1$) suggesting that Look-ahead users were still increasing their efficiency at planning moves by trial 6.

A trend analysis on 1-Move performance across trials revealed both significant linear, $F(1, 18) = 17.39$, $p < .001$, $MSE = .57$, and quadratic curve components, $F(1, 18) = 26.26$, $p < .001$, $MSE = .08$.

Look-ahead Span

As previously described Look-ahead interface users could enter 3 or more moves at a time. Table 1 below shows the average number of moves entered across trials.

Table 1

Number of moves entered per return key press for Look-ahead interface users

Trial	Mean Look-ahead	Standard Deviation
1	3.04	0.17
2	2.98	0.17
3	3.09	0.38
4	3.11	0.65
5	3.01	0.14
6	3.06	0.29

It appears from the results in Table 1, that specifying a 3-move sequence was perhaps the limit for most participants. None appear to have taken advantage of being offered the opportunity to enter more than 3-moves, again suggesting that 3-moves at a time was a difficult task for most participants.

Discussion

Contrary to expectations, the initial attempt to increase planning and look-ahead by forcing the specification of a small number of steps failed to result in an observed increase in performance. The effect of a look-ahead interface did not significantly increase performance in terms of number of moves or a reduction in the proportion of palindromic moves made. The results indicated that while there was no overall effect of interface on move performance, the extra planning did not come at the expense of total solution time with both 1-Move and Look-ahead users completing trials in approximately the same amount of time.

From the results of trial 1 performance, participants using the look-ahead interface appear to have been somewhat hampered in their first attempts to solve the 8-puzzle, with a dramatic increase in both time and number of moves being evident. Although the look-ahead interface appears to have made subsequent performance less variable following trial 1, with users appearing to adapt to the interface requirements of specifying and moving at least 3-tiles at a time, they were seemingly unable to build upon their increased experience with the puzzle and perform more efficiently by the end of trials.

Inter-move latencies, although shorter by the end of the six trials, also remained significantly more than the control group over trials suggesting that participants were still having to think extensively about their move selections while 1-Move latency times became significantly shorter as trials progressed. This is an indication that the refinement in performance evidenced by O'Hara & Payne's (1998) planning group was not being tapped into in the same extent by the current manipulation.

The expectations of observing differences in the number of palindromic sequences made by each group were not evident either which may either simply be a symptom of the lack of performance benefits that the current manipulation failed to induce or simply because the current manipulation does not have the appropriate interactive mechanisms in place that would discourage backtracking as previous cost approaches have successfully done (Delaney & Knowles, 2005).

A high-cost operator interface would by its nature reduced the amount of back tracking as it is expensive to subjects in terms of the time and effort to undo a move or a sequence of moves. The same cost however is not associated with either interface in the current experiment. Undoing 3-moves is a simple process for look-ahead interface users, taking little time or effort to undo a previous selection. This would be especially true if the moves were still fresh in memory and if the resulting state was immediately evaluated as unproductive for reaching a solution. With the lack of a cost quota linked to poor move selection it would then appear more unlikely that any evidence of differences in the ratio of backtracking moves should be found. From the palindrome ratio performance data in Figure 6 it actually appears that the look-ahead condition had a higher proportion of palindrome sequences on average across trials. This again may be an indication of the lack of clear planning induced by the current

look-ahead manipulation. Alternatively, the interface may have increased exploration from a given point which would be assessed and if judged unfavourable users would then return to their previous state to select a different sequence of moves.

Despite the lack of performance differences in terms of moves and latencies, interface did not appear to significantly impact upon the time taken to complete puzzles. Apart from an increase in time taken to solve trial 1, no other trials appeared to differ significantly in their solution times. It would appear that increasing look-ahead did overall not have a negative impact for participants.

The current experiments may cast some doubt on the conclusions of O'Hara & Payne (1998) about increased look-ahead and planning being induced by a high-cost operator interface. It may have simply been that users were selecting their next move more carefully from the available successors and that the mechanism responsible was simply an increase in the quality of their choices about possible appropriate successors. As Jefferies et al. (1977) suggested in their model of water jars performance, problem solving behaviour may be adequately predicted by having a look-ahead of only1 step. In terms of the previous finding therefore the cost interfaces may have simply shifted the balance towards considering each move more fully in a breadth first manner as opposed to depth of search.

There are however a number of reasons that suggest such a conclusion may be premature. Firstly, look-ahead performance in the current experiments bares no resemblance to the performance observed by O'Hara & Payne (1998). The current results indicate that whatever mechanisms were responsible for the successful benefits

gained from implementing a high-cost operator interface it appears those same mechanisms were not being activated in the current study. Although caution is warranted when making such comparisons with the previous study, the lack of any similarities is quite striking. Look-ahead performance by trial 6 was 76 moves on average while in the previously reported studies participants were nearing the optimal performance of 17 moves by the same trial.

Secondly, latency times were not being reduced to the same degree in the current experiment as in previous experiments. Whatever mechanism was responsible for decreasing the time needed to consider moves previously were once again not being induced in the current manipulation.

Previous research has suggested that only around 33% of participants will typically look-ahead when solving a problem. If this is the case the current interface may have excluded almost two thirds of the population who are generally uncomfortable with multi-step look-ahead sequences. Recent research on TOH isomorphs (Gunzelmann & Anderson, 2003) has found that people naturally increase planning with practice. Perhaps the lack of an initial phase in which to become accustomed to the task may have been detrimental for many of the look-ahead interface participants. The poor performance on trial 1 suggests that this may have been the case. A second argument that Gunzelmann & Anderson (2003) make and which the current argument also supports, is that planning will occur if participants can see the *utility* of increased levels of planning as it may often lead to better performance and allowed the completion of trials in a more timely fashion. This leads to an interesting suggestion that motivation, real or perceived, could be an important factor in the successful

implementation of increased planning interfaces. Take for example high-cost interface manipulations that have been found to increase performance. The benefits of increased planning would have been inherently obvious to the users of those interfaces. By not planning, greater effort would be required by participants to enter moves and this effort would have to be maintained until the task was completed. By becoming more efficient at solving the puzzle the utility of increased planning would have become obvious to all participants. The current manipulation however asked participants to simply do more work yet none of the hoped benefits seemingly transpired. Many of the participants commented that it took a lot of mental effort to think 3 moves and they could not understand the reasons for such a requirement. Previous manipulations would not have had to contend with such a point as participants would themselves decide to plan more at their own rate and would feel the benefits in terms of solving trials in smaller numbers of moves.

It appears that strict enforcing of look-ahead may not be an avenue that can be explored, at least with users of a new interface and on a novel problem. Dictating performance in a set number of mental units appears to lead to performance and solutions that are average at best. Expert performance in a task has often been described as creating and proceduralising a number of units of behaviour or mental sequences into productions that reduce the need to plan individual sequences (e.g. Anderson, 1990; 1993). However, the composition and size of these chunks may vary greatly from participant to participant and over trials. The 3-move interface used in experiment 1 would not have supported the individually self paced development of these planning units. Therefore, encouraging look-ahead from the mental unit of one step upwards rather than attempting to strictly enforce it from a pre-determined

limited set may produce the sought after performance benefits. The measuring of these mental units over trials may also allow a much fuller description of the increase in look-ahead while problem solving. This issue was explored more thoroughly in Experiment 2.

Experiment 2

The strict enforcement of entering a specific number of moves at once did not automatically lead to the benefits in performance that were hoped for. It appears from experiment 1 that 3-moves may either be simply too difficult for participants, or that an interface aimed at increasing look-ahead needs to reflect the varied nature of planning by being as flexible as possible.

Secondly, a motivational mechanism to highlight the need for increased planning, whether it is artificial in terms of its implementation or intrinsic to the interface may also be needed if participants are to increase their efforts beyond the typical minimum. An encouragement, as opposed to enforcement, of look-ahead in aid of some cause or higher goal may provide the ideal manipulation. If there was a perceived purpose for the increased planning then participants may be more likely to partake in it.

The current experiment aimed to increase performance by providing feedback of performance to participants through the introduction of a 'Scoreboard' (see Figure 9) that contains a score for each trial based on the number of times the 'Return' key is pressed. The aim, as in the previous experiment, was to solve the puzzles in the fewest number of moves possible although participants previously had had no external feedback as to how well they were performing. In the current manipulations they are not only able to see how well they are performing on a particular trial but also to compare their performance over trials with the aim of improving as trials progress.

Method

Participants

61 undergraduates aged between 18 to 27 years (Mean 20.35 years, S.D. = 2.35) took part and were paid £6 or given course credit. None of the participants claimed to have had any experience with the 8-puzzle although it was generally known by participants. 21 participants, ten from the 1-Move condition and eleven from the look-ahead condition, were unable to complete the trials within the allotted time and therefore were excluded from the main analyses but are discussed in the results section.

Design

Similar to experiment 1, the current experiment was a 3-factor mixed design with the within subject factor of trial having 10 levels. The between subject factor of problem start state, with 2 levels, are identical to the start states used in Experiment 1 (see Figures 2a & 2b). Finally, the between subject factor consisted of two levels of interface; a control 1-Move condition and a self paced 'Look-ahead' group that could enter any number of moves over and above the required 1 move specified entirely at the users own discretion.

The 8-puzzle program once again recorded total time, number of moves and latency times for both groups during all trials. Number of moves minus palindromes and ratio of palindromes were also calculated for both groups. For the look-ahead group a measure of the extent to which participants were looking ahead was calculated by dividing the total number of moves entered by the number of times the 'Return' key

was pressed (i.e. the score for a trial). This measure can be used to provide a rough estimation of the extent to which participants were mentally searching along the solution path.

Materials

The puzzles appearance and the keyboard controls specifically used to manipulate tiles in the 8-puzzle were identical to those used in experiment 1.

The interaction for 1-Move subjects was also the same as in experiment 1. Participants would simply enter the tile they wished to move into the command line and press the 'Return' key, upon its validation the tile would move into the intended space and the score for that particular trial would increase by 1 on the scoreboard which was located immediately to the right of the 8-puzzle (Figure 9 below).

Look-ahead interface users were also required to enter a move(s) and to press the 'Return' key that would also increase their score by a value of 1. However, users could enter a tile sequence of any length and upon its validation all tiles specified from the compulsory 1 tile to the legal maximum number entered would then be moved and the score for that trial would still only increment by a value of 1. Therefore, greater numbers of tiles can be moved yet still only increase a score by a small amount.

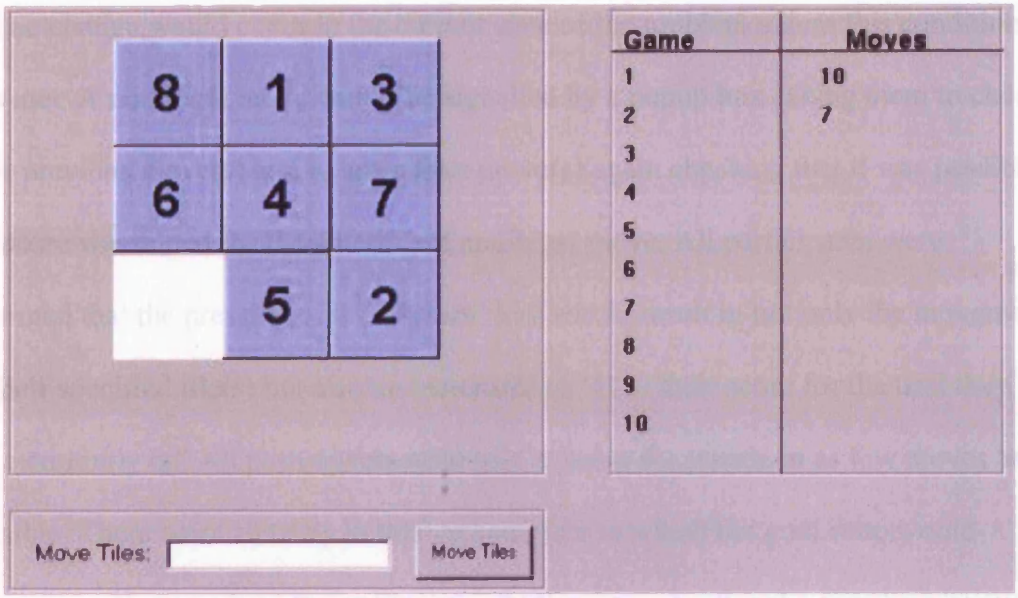


Figure 9. The Scoreboard Interface used in Experiment 2

For all participants if an illegal move was entered a message box would appear informing them of the problem, the command line would be cleared and participants would be left to re-enter their move(s) once again, fixing whatever problem had occurred previously.

Procedure

Participants were informed that they were taking part in a problem solving experiment. Seated in front of a computer they were shown an example of a typical start state (not used in the actual experiment) and told the aim was to transform the start state into the goal state. All participants that only keys 1-8 could be used to control the movement of the tiles along with the 'Return' key which would then implement their chosen move. For users in the look-ahead condition they were informed that more than 1 move could be implemented at once. They could specify a sequence on moves of any length from the minimum of 1 to a sequence of any length of their choosing. Participants were informed that any move(s) entered must be legal

and no change would occur in the current state of the problem unless this condition was met. A non-legal move would be signalled by a popup box asking them to check their previous move(s) and to enter their move(s) again checking that it was possible. No score was added for the entering of an illegal move. All participants were informed that the pressing of the 'Return' key would result in not only the movement of their specified tile(s) but also an increment of '1' to their score for the trial they were currently on. All participants were told to solve the puzzle in as few moves as possible. There were 10 trials in total to complete in which the goal state would remain constant as would their start state. Participants completed a simple rotation of 6 tiles to confirm that they understood the controls and were told to press the button marked 'Begin' when ready. Participants were informed of a successful completion of a trial by a popup box and were told to press the 'Begin' button again to start the next trial.

Results

Unfinished Groups

As mentioned previously a total of 21 participants could not complete all 10 trials. However, this was a consequence of being unable to either solve even a small number of the puzzles before asking to be excused from the experiment.

It is unlikely that differences in ability to solve the puzzle are a consequence of either age or population sample differences. The almost equal numbers failing to solve the puzzle would also indicate that it is not a problem with a particular interface but rather

due to a fundamental inability to reach a solution(s) by a particular subset of participants.

Identifying one particular reason for this apparent difficulty is complicated by the sheer number of possible states that the 8-puzzle can occupy estimated by 9!/2, giving 181, 440 possible states. Attempting to find or identify a singular or particular pattern reason for the difficulty, or what is more likely a number of sources of difficulty, is very problematic. The best approach therefore may be to characterize the performance data of participants who were unable to complete the trials, examining in particular the trials that were eventually abandoned or trials that took exorbitantly large numbers of moves to complete.

To begin teasing apart some of the possible reasons a comparison of performance on the first three trials for total time, total moves and inter-move latency times are shown below in Figures 10 to 12. Caution must be urged however as some trials from these points onward were left unfinished. Participant SR in the 1-Move Unfinished group was excused from the experiment after making 486 moves on trial 2, three participants from the 1-Move group and one participant from the look-ahead group could not complete trial 3 and asked to be excused from the experiment. Therefore trial 3 may be a more conservative estimate of performance than would have been expected if they had gone on to actually complete the trial. The inclusion of trial three performance measures however reveal an important aspect to the unfinished group's performance and warrant inclusion, although caution is prescribed on any conclusions that may be drawn.

Figure 10 below, compares the total time to solution for 1-Move groups who completed trials versus those who did not as well as look-ahead users who also completed all trial versus those who did not.

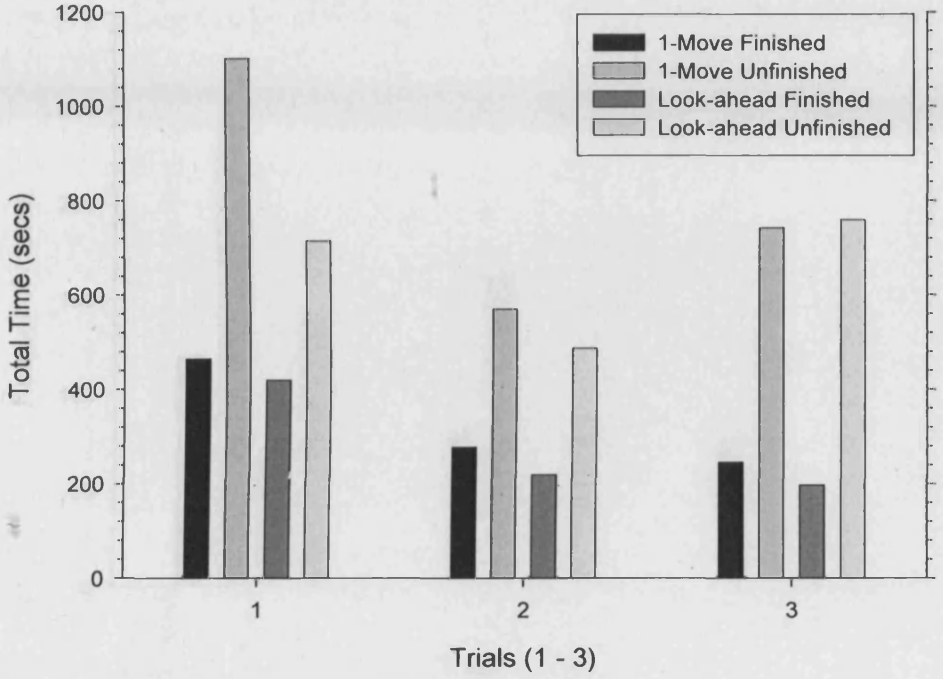


Figure 10. Total Time to Solution for both Interface Groups for Finished and Unfinished Groups

The graph above demonstrates two important points. Those in the unfinished group are seemingly quantitatively different from those who managed to complete all trials with seemingly little trouble. Secondly, the unfinished groups appear to not learn in the same way across trials compared to those who finished all trials. However caution is needed when making such judgements due to missing values for unfinished groups.

Examining the total moves made by participants of both interface groups and for both completion and non-completion participants reveal similar patterns to the total time data.

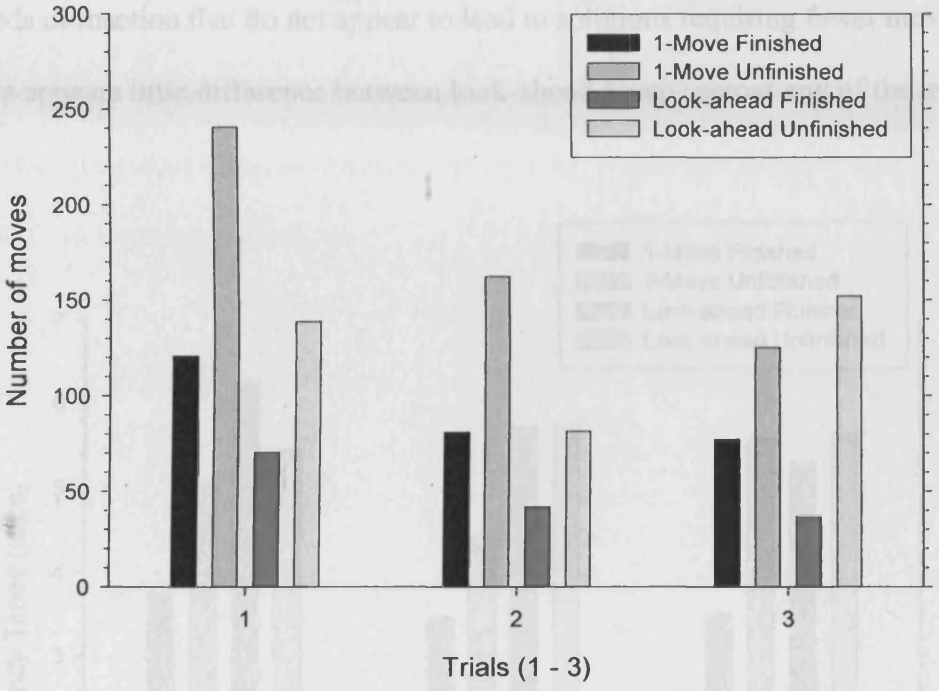


Figure 11. Total Moves to Solution by both Interface Groups for Finished and Unfinished Groups

The total move data also reveal that the unfinished groups are different from those who completed all trials. The pattern seems to suggest large total move differences by the 1-Move unfinished group across all trials and rather erratic non-linear performance by the look-ahead unfinished groups. The 1-Move unfinished groups trial 3 total move performance may also have demonstrated the non-linear performance if the three participants who stopped without reaching a solution had actually gone on to complete the trial.

The inter-move latency times for all groups are shown below in Figure 12. 1-Move unfinished participants appear to not share the same inter-move latency pattern as those in their respective finished group. On trial 3, there is an increase in inter-move latency suggesting that difficulty is perhaps being experienced leading to longer periods of inaction that do not appear to lead to solutions requiring fewer moves. There appears little difference between look-ahead groups across any of the trials.

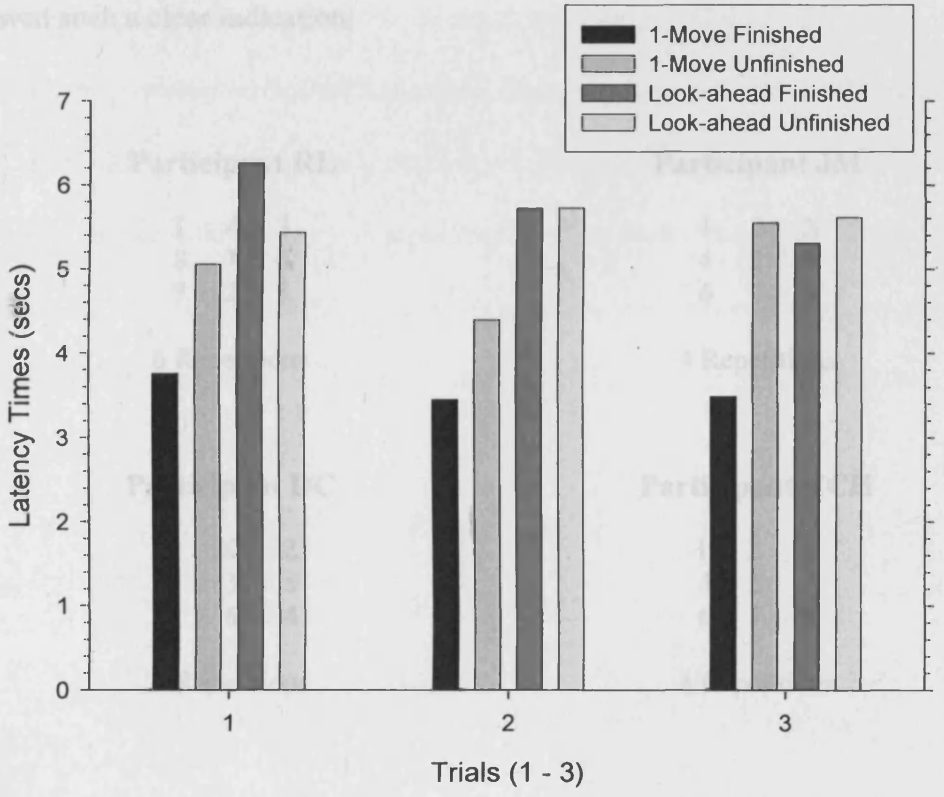


Figure 12. Inter-Move Latency Times for Trials 1 to 3 for Finished and Unfinished Groups

The large number of possible states with which the 8-puzzle can enter makes identifying one or more key problem states very problematic. Rather than the source of difficulty being one particular state it would be more likely that there are in fact several sources of difficulty in the 8-puzzle. The ability to satisfy multiple subgoals

simultaneously would appear to be the most likely candidate for problem solving difficulty. The exact source is difficult to categorise but from the general observations the attainment of multiple subgoals appears for some participants a difficult task. For example the participant HB on trial 2 made a total of 454 moves before giving up. Not once in the entire attempted solution did HB attain the '123' formation in the 8-puzzle. This would indicate that the primary source of difficulty for HB was the acquisition of this particular arrangement of tiles. Not all participants' data however showed such a clear indication.

Participant RL

1 4 3
8 X 6
7 2 5

6 Repetitions

Participant JM

1 2 X
4 3 5
6 7 8

4 Repetitions

Participant DC

1 X 2
8 3 5
7 6 4

5 Repetitions

1 2 X
8 3 5
7 6 4

4 Repetitions

1 3 2
8 5 X
7 6 4

4 Repetitions

Participant CCH

1 2 X
4 3 5
6 7 8

4 Repetitions

1 X 2
4 3 5
6 7 8

5 Repetitions

Figure 13. Frequent Repetitions of Problem States by Unfinished Participants

There are however other arrangements that appear to cause equal difficulty once the '123' arrangement of tiles has been resolved. Figure 14 below shows that participants can have equal difficulty with other competing arrangements that lie anywhere in the problem space.

Some participants did in fact achieve the '123' subgoal early in the problem yet still went on to make large amounts of moves to solve the puzzle, indicating the problems they were experiencing were involved with either the attainment of the '56', '78' subgoals or attaining both simultaneously. For example, participant SR reached the '123' subgoal within 31 moves on trial 1 yet finished the puzzle with 307 moves in total, 276 moves after one key subgoal had been attained. In contrast to this level 1 performance, SR then took 486 moves to reach the '123' arrangement on level 2 indicating that the source of difficulty had shifted to another particular subgoal or set of subgoals.

Participant AJ

1	2	3
5	6	4
8	X	7

3 Repetitions

1	2	3
5	6	4
X	8	7

2 Repetitions

Participant SR

1	2	3
7	X	5
8	6	4

4 Repetitions

Participant PJ

1	2	3
6	8	5
7	X	4

3 Repetitions

Figure 14. Other sources of difficulty in arranging tiles

The one defining characteristic of all the non-solvers, irrespective of interface, is the apparent lack of consistency in their performance – indicating quite possibly a lack of learning about the problem. Trials occurring much later in the solution can also be as likely to lead to problem solving failure than trials encountered early in the process. For example participant PJ made a total of 321 moves on trial 5 after having solved the previous two trials with 61 and 65 moves respectively. It appears then that difficulty can occur at any time during problem solving irrespective of the number of trials having being solved previously. It may be that the 8-puzzle requires a reasonable period of exposure in terms of practice and familiarity before consistent performance is observed in terms of being able to complete a puzzle.

Main Data Analysis

The 40 participants who actually completed all ten trials are now presented. Data were log transformed to stabilise for variance and were analysed using a 3-way mixed ANOVA.

Total Moves

The effects of interface on total number of moves made can be seen below in Figure 15.

Main analysis revealed, in line with predictions, a significant effect of interface, $F(1, 36) = 14.59$, $p < .01$, $MSE = 4.77$, with those in the look-ahead manipulation solving the problem in fewer moves across trials. The analysis also showed a significant effect of trial, $F(9, 324) = 19.01$, $p < .001$, $MSE = .78$.

There were no significant interactions between trial and interface, trial and problem type or a trial x interface x problem type interaction (all F's < 1).

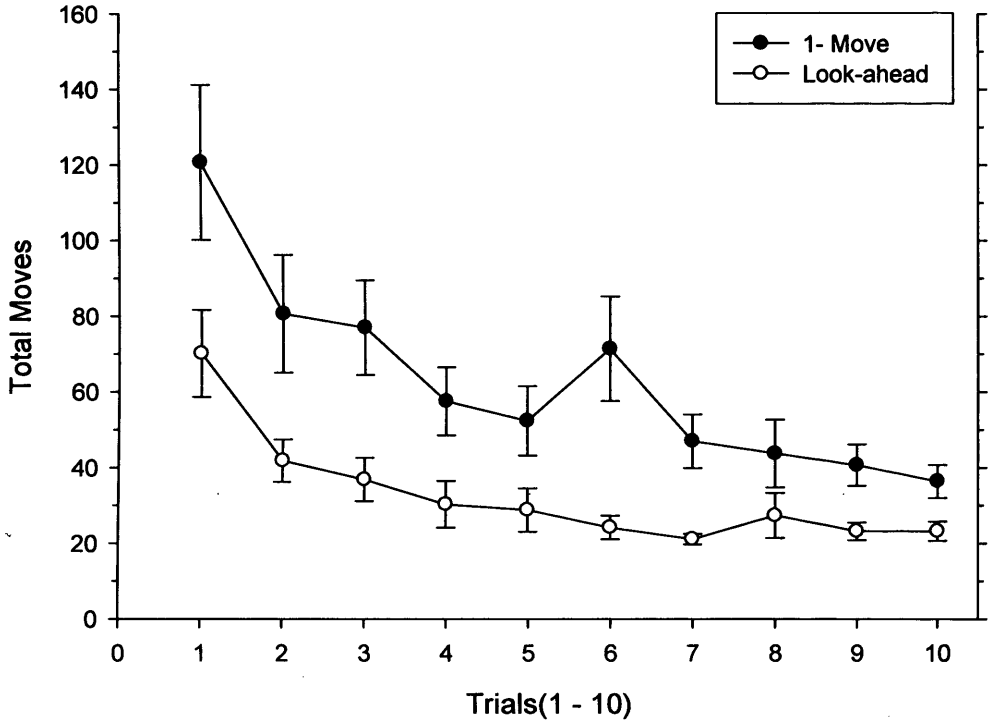


Figure 15. The Effect of Interface on Total Moves to Solution across Trials

Analysis of the trends found a significant linear trend for 1-Move interface users, $F(1, 19) = 26.63$, $p < .001$, $MSE = 2.99$. There was no quadratic component to the trend, $F(1, 19) = 1.98$, $p > .1$, $MSE = .19$.

A significant linear component was also present for users of the look-ahead interface, $F(1, 19) = 62.98$, $p < .001$, $MSE = 2.56$. Unlike the 1-Move group there was also a significant quadratic component to the curve, $F(1, 19) = 26.62$, $p < .001$, $MSE = 1.04$.

Total Moves minus Palindromes

There was as significant effect of interface with those in the look-ahead condition making significantly less moves, $F(1, 36) = 14.961$, $p < .001$, $MSE = 4.364$.

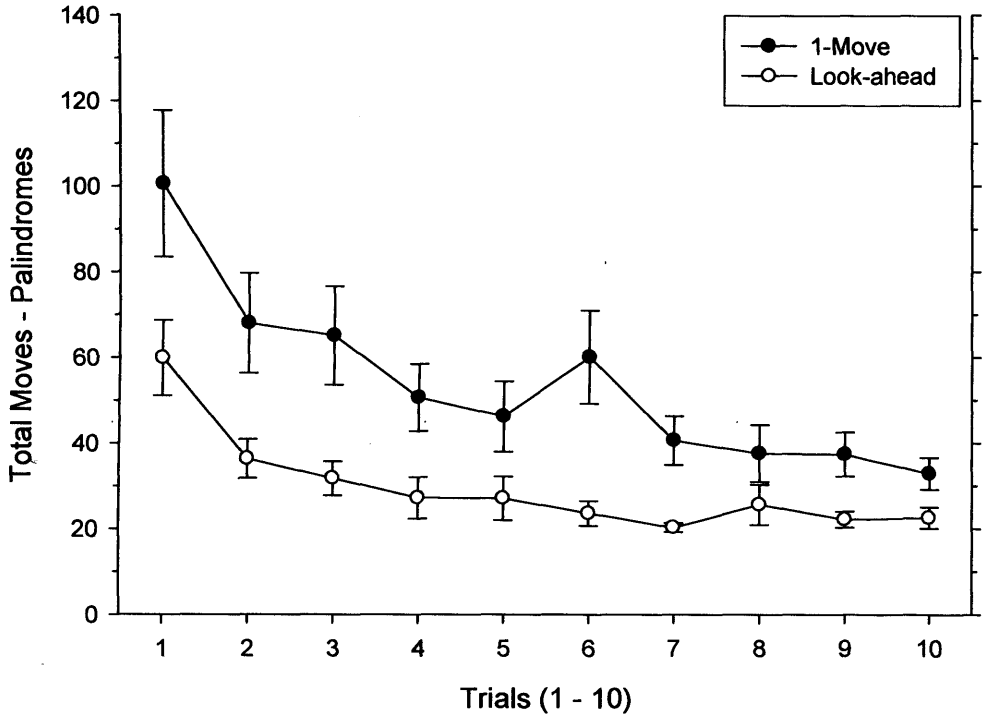


Figure 16. Total Moves - Palindromes

There was a significant effect of trial, $F(9, 324) = 17.727$, $p < .001$, $MSE = .63$. There was no significant effect of start configuration or interactions.

Contrasts revealed a significant linear trend for the 1-Move group, $F(1, 18) = 24.79$, $p < .001$, $MSE = 2.57$. There was no quadratic component to the curve, $F(1, 18) = 2.27$, $p > .1$, $MSE = .182$.

Contrasts revealed a strong linear component to the curve, $F(1, 18) = 48.19$, $p < .001$, MSE = 1.82. There was also a significant quadratic component to the curve, $F(1, 18) = 23.64$, $p < .001$, MSE = .82.

Palindrome Ratio

The main analysis revealed no effect of interface, $F(1, 36) = 2.572$, $p > .1$, MSE = .036, on the ratio of palindrome moves to total moves.

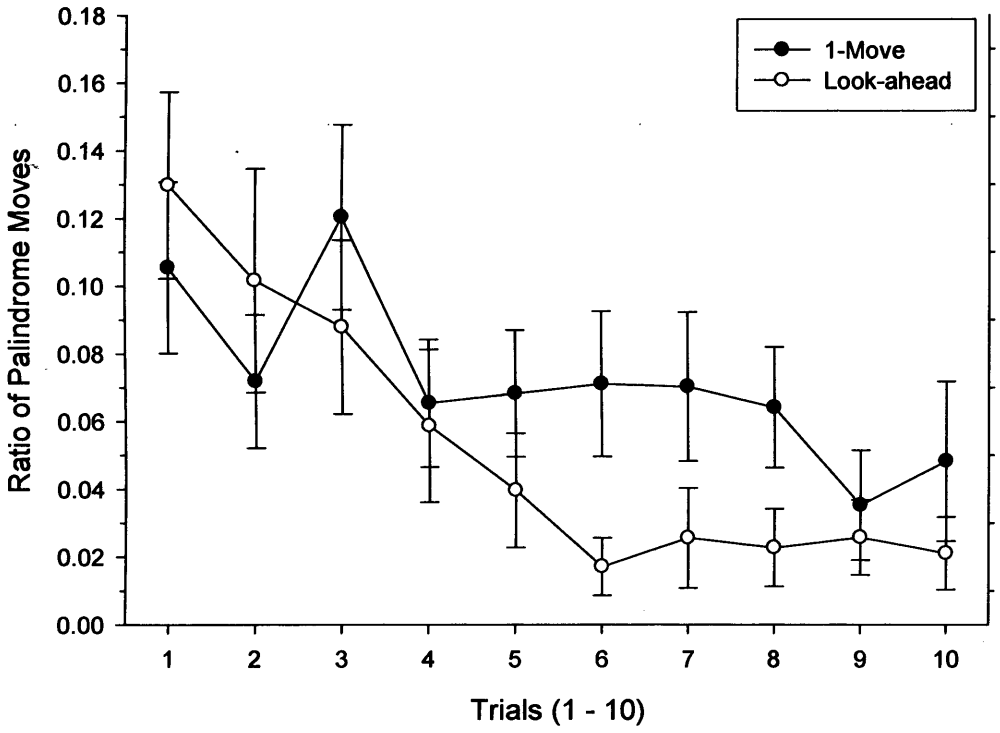


Figure 17. Ratio of Palindrome Moves

There was a significant effect of trial with less Palindromic moves sequences being made as experience increased over trials, $F(9, 324) = 4.69$, $p < .001$, MSE = .04. There were no significant interactions or other main effects.

Trend analysis for ratio data showed that for the 1-Move group there is evidence of a linear component, $F(1, 18) = 4.392$, $p < .051$, $MSE = .065$, but no quadratic component ($F < 1$).

Trend analysis revealed a significant linear component for the look-ahead condition, $F(1, 18) = 24.56$, $p < .001$, $MSE = .23$, and a significant quadratic component, $F(1, 18) = 6.75$, $p < .02$, $MSE = .05$.

Total Time

The effect of interface upon total time to solutions across trials can be seen below in Figure 18.

Analysis of the total time to solution revealed no effect of interface on time taken to complete trials, $F(1, 36) = 1.59$, $p > .1$, $MSE = .89$. There was also a highly significant effect of trial, $F(9, 324) = 34.16$, $p < .001$, $MSE = 2.66$. There were no significant interactions at any level between trial and interface, trial and problem type or a trial x interface x problem type interaction (All F's < 1).

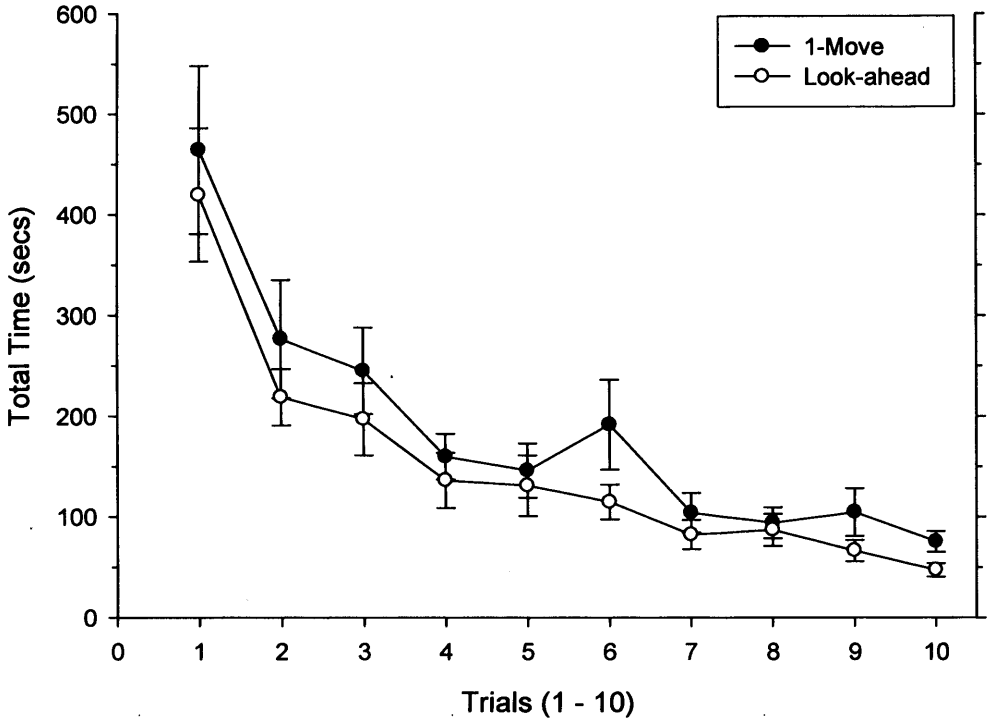


Figure 18. The Effect of Interface upon Time Taken to Solve Problems

For the 1-Move group a trend analysis revealed a strong linear component, $F(1, 19) = 40.99$, $p < .001$, $MSE = 9.22$. The trend analysis indicated a small quadratic component to the curve but did not quite reach significance, $F(1, 19) = 3.04$, $p > .05$, $MSE = .37$.

For the look-ahead group, analysis of the trends revealed a significant linear component to the curve, $F(1, 19) = 97.04$, $p < .001$, $MSE = 13.20$. Similar to the 1-Move condition there was also an indication of a quadratic component but this also did not reach significance, $F(1, 19) = 2.66$, $p > .05$, $MSE = .537$.

Inter-move latency

The effect of interface upon inter-move latencies across trials are shown in Figure 19 below.

There was a significant effect of interface, $F(1, 36) = 5.45$, $p < .05$, $MSE = 1.54$. The effect of trial was highly significant, $F(9, 324) = 19.44$, $p < .001$, $MSE = .73$, and the effects were moderated by a small but significant trial x interface interaction, $F(9, 324) = 2.23$, $p < .05$, $MSE = .08$.

Simple main effects analysis revealed a significant difference between the groups' latency times at trial 1, $F(1, 36) = 10.23$, $p < .01$, $MSE = .29$, trial 2, $F(1, 36) = 10.819$, $p < .01$, $MSE = .36$, and trial 3, $F(1, 36) = 7.12$, $p < .05$, $MSE = .29$. The difference between groups disappeared at trials 4 and 5 yet reappeared at trial 6, $F(1, 36) = 8.77$, $p < .01$, $MSE = .56$. Although trial 7 approached significance it did not quite reach significance, $F(1, 36) = 3.40$, $p < .1$, $MSE = .244$. The remaining three trials revealed no significant differences in latency times between interface groups (all p's $> .1$).

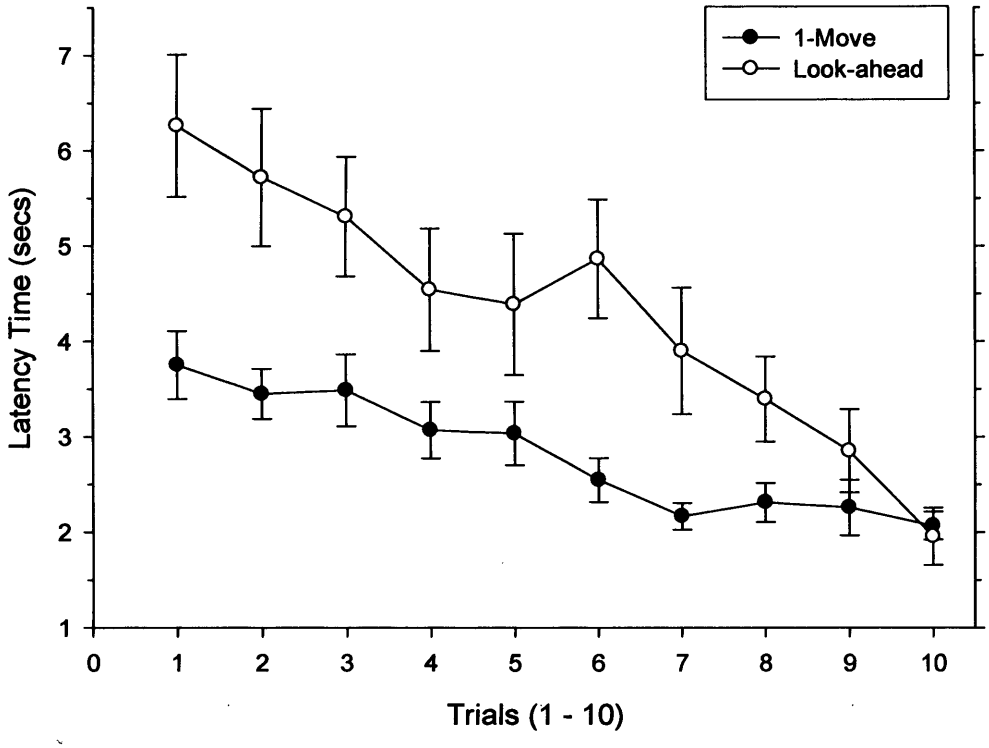


Figure 19. The Effect of Interface on Latency Times across Trials

The analysis of trends revealed a significant linear component for the 1-Move group, $F(1, 19) = 35.08$, $p < .001$, $MSE = 2.00$. There was no quadratic component to the curve, ($F < 1$).

For the look-ahead condition there was also a significant linear component, $F(1, 19) = 32.09$, $p < .001$, $MSE = 4.67$. Similarly to the 1-Move group, there was also no evidence of a quadratic component to the curve, $F(1, 19) = 1.14$, $p > .1$, $MSE = .18$.

Look-ahead Measurement

As previously stated the current experiment allowed a measurement of the extent to which participants were looking-ahead during their problem solving

attempts. Figure 20 below shows the number of tiles moved on average at once as specified by participants.

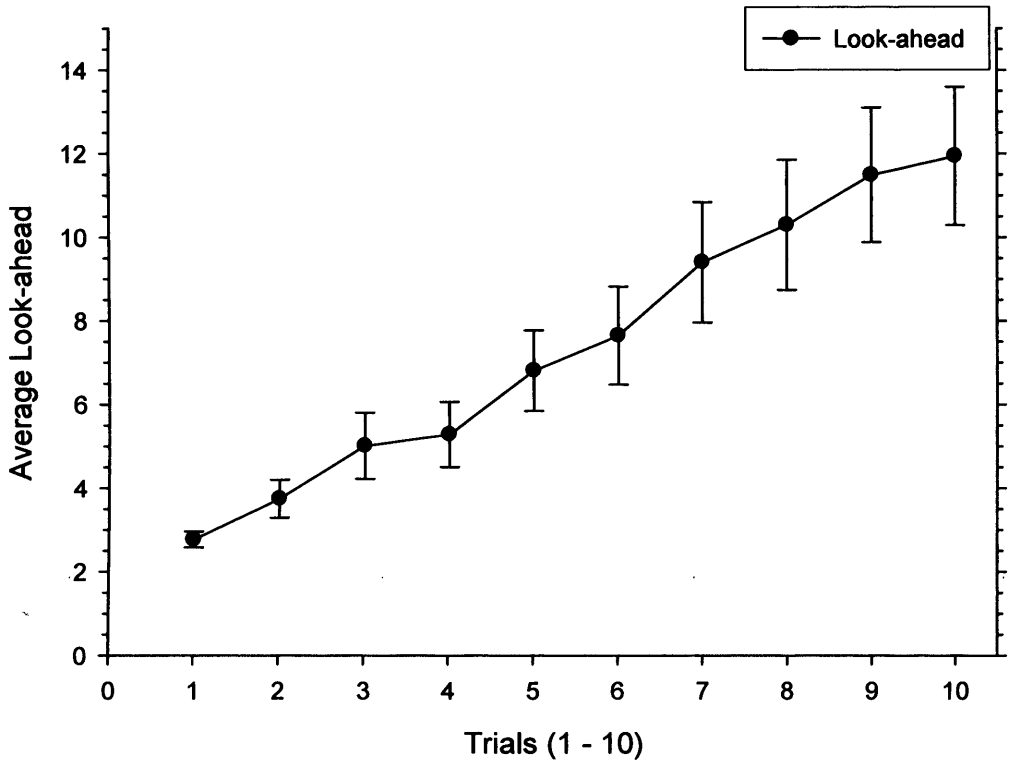


Figure 20. Average Number of Tiles Entered at Once by Look-ahead Interface Users across Trials

A within subject ANOVA with start state as a between subject factor revealed a highly significant effect of trial, $F(9, 162) = 28.023$, $p < .0001$, $MSE = .70$. There was no effect of start state or an interaction (all F's < 1).

A within subject trend analysis on look-ahead values over trials revealed a highly significant linear component to the trend, $F(1, 19) = 47.62$, $p < .001$, $MSE = 6.13$. There was also evidence of a quadratic component to the trend, $F(1, 19) = 4.917$, $p < .05$, $MSE = .15$.

Discussion

The current experiment's attempt to replicate previous findings of increased 8-puzzle performance while providing clear evidence of the mechanisms involved being greater planning and look-ahead were successful. The results stand in contrast to those of experiment one's, whereby the manipulation attempted to increase performance by enforcing a strict number of pre-planned moves to be specified. By the end of trials in the previous experiment, performance appeared to still resemble that of unskilled problem solvers and was not significantly different from controls. From the performance observed from the final trials in the current experiment, problem solving appeared to be almost optimal in its efficiency with both total time and inter-move latency times being indistinguishable from the normally rapid performance of 1-Move controls.

From the look-ahead values obtained from trial 1 performance it appears that on average a value of just over two steps appears to be the preferred depth of look-ahead search when solving the puzzle under a largely self paced planning manipulation. Whilst look-ahead can increase with greater experience it seems that during the initial period of becoming accustomed with the task, depth of search remains relatively small and requires large periods of planning before any action is decided upon and moves are actually implemented. The noticeably poorer performance of look-ahead participants from trial one of the previous experiment may be somewhat explained by the average two step look-ahead that appears preferred for initial performance in the current experiment.

Overall, the performance data in terms of number of moves, total time and inter-move latencies is equivalent to that of previous successful cost based manipulations (O'Hara & Payne, 1998). It suggests that the proposed planning and look-ahead mechanisms invoked by those manipulations and responsible for the increase in performance are also currently being invoked by the present scoreboard manipulation.

The current manipulation appears to have tapped into the more opportunistic style of planning that may typify human planning behaviour (Hayes-Roth & Hayes-Roth, 1979). The rigidity of the previous interface manipulation would appear to have been at least a major reason for its failure to increase problem performance. With the current interface, there appears to have been a larger adaptation to the task with increased levels of planning occurring when the optimal solution path has been discovered. From the initial trials there appears to be a gradual refinement of the solution path with more efficient solutions being generated with each new attempt. The final trials remain almost constant in their solution lengths, suggesting that an optimal or near perfect solution had been discovered and look-ahead was then increased with larger numbers of tiles being entered at once to further improve the participants 'move score', rather than the solution path itself.

While the large number of moves entered by the end of trials appears to be counterintuitive to most theories of working memory it appears that a chunking of large sequences of moves was evidenced in the final few trials which would account for the large increases in number of moves specified. If as the data suggest, the optimal solution path had already been identified in earlier trials and move choice had become increasingly refined, it would suggest there was a large memory component

for particular sequences of trials. Inter-move latency times, by the end of trials were no longer significantly different, as large numbers of trials were being entered at once and in a rapid fashion indicative of final behaviour often observed during the concluding phase of a problem trial, generally when the solution becomes known to participants (Kotovsky et al., 1990).

In line with previous planning manipulations no difference in total time to solution appears to result as a consequence of the increased planning due to the generation of more efficient solutions. Although, compared to the previous manipulation, 1-Move control subjects also appear to have become more motivated to increase their levels of performance. This increase is most likely due to the effect of visual feedback on their trial performance and may have increased their levels of planning above the norm. However, performance can still be extended by the introduction of an additional look-ahead manipulation and increase planning further still.

However, a number of issues still remain, the most important of these centring around the large number of non-finishing participants taking part. Other important questions exist such as how behaviour and performance react when the start state is not a constant factor. The large number of moves being entered in the current experiment, it has been argued, is most likely due to the formation of increased numbers of chunks of moves that enable such depth of look-ahead to occur. When such constant sequences are not actually available over repeated trials, would look-ahead still increase in such a linear fashion? These questions in particular form the basis for Experiment 3.

Experiment 3

While the second experiment appeared to tap into the mechanisms responsible for increased performance found by O'Hara & Payne (1998) the results were complicated by several findings. The motivation for carrying out the current experiment therefore centred around two main factors in particular.

Firstly, an effort would need to be made in finding a method(s) of reducing the number of participants not completing trials. One particular method often used in problem solving research has been the introduction of a hint in an attempt to aid performance (e.g. Gick & Holyoak, 1983; Weisberg & Alba, 1981a). If as the current suggestion has been, unsuccessful performance may centre around being unable to accomplish the simultaneous ordering of particular tiles, due perhaps in some part to self imposed constraint(s) about the problem. It may follow that increasing awareness of such constraints will hopefully alleviate the problem, for what appears to be only a particular subset of participants. Ericsson (1975) classified from participants' protocols a typically weak model of 8-puzzle performance called the 'Single Tile Difference Model' in which a hill-climbing strategy of focusing upon placing a single tile in its position before moving on to the next tile was a large motivation behind much of the observed performance. While this is not an unusual behaviour, it suggests a limited attention to the problem configuration in relation to the overall goal configuration. Ericsson (1975) also describes another model of performance which he terms the 'Distributed Attention Model' of 8-puzzle performance as an indication of participants who were more successful 8-puzzle solvers. The model has at its foundation a classification of behaviour that is indicative of considering not single or even pairs of tiles but viewing the 8-puzzle as an entire sequence of linked digits, that

if considered as a chain would allow for greater performance. The look-ahead interface may induce this type of behaviour yet may not naturally induce it amongst all participants. As Ericsson (1975) states, "Whereas in the earlier methods the selected intention determined the sequence of moves to be made, the opposite is true for this method. Here the alternative moves or sequences of moves are explored and evaluated with respect to the features of the GC that can be attained, and in a sense generate the intention" (Ericsson, 1975, p. 68). Therefore, relations between multiples tiles are considered from the very beginning and the attainment of a single subgoal does not receive all user attention at the expensive of the wider consideration of the goal configuration. Successful participants may not confine themselves to placing a single tile in its place and also are not constrained by unmoving tiles once correctly placed. It suggests therefore that increasing all participants' awareness of such features may help increase the overall number of successful completions.

As an additional intervention, a time limit may also prove a valuable tool that allows the exiting of a puzzle if after a sustained period of time attempting to solve, no solution has yet been reached. Allowing such a manipulation whereby participants have the opportunity to at least attempt all trials will give a much clearer indication of the success of the current manipulations in terms of overall performance and possibly even learning. If participants are unable to complete large numbers of trials then the manipulations success as a means for improving performance and learning, would be under doubt. If however, overall pass rate remains high, it would suggest that the previous findings simply showed, as the current argument has been, that it was an inability to complete one or a set of subgoals at particular stages in the process that was responsible for the failure, due to a poor conceptualisation of the problem task.

An effort to find out if the failure rate is really as high as Experiment 2 suggests or whether it is simply a lack of conceptual knowledge that can easily be rectified through a hint intervention should provide some answers. Another possibility is to examine how solution length is related to pass rates. Shortening the solution path by almost 1/3 should lead to problems that are fundamentally easier. If no differences in the solution rates between lengthier problems and short problems then there appears to be more evidence that the problem is dependent on participants solving particular subgoals and that the interface can be used with problems of increased and decreased complexity. If there are no significant differences in rate of solution success in relation to solution path length, then this factor can be eliminated as the main source of problem difficulty.

The second aim was to provide both a replication and a refinement of the Look-ahead interface groups' results from experiment 2. The apparently large amount of look-ahead performance exhibited by the look-ahead interface users suggests that additional mechanisms such as memory for particular sequences, in terms of chunking large numbers of moves rather than problem solving per se, may have been operating over trials for look-ahead users. Would the observed linear increase in number of tiles entered (i.e. average look-ahead) over trials still be observed when the start state does not remain constant? If the problem state changes over trials then there may be the observation of a much more consistent number of tiles entered per return key press as opposed to the linear increase previously observed. If the increase does still exist it leads to an interesting suggestion that look-ahead may be a component process akin to a skill like many other problem solving mechanism, which develop with increased exposure to a particular area. Furthermore, the ability to outperform controls

consistently across trials would also suggest that this is a transferable skill rather than simply benefiting performance on repeated exposure to one problem state in particular.

Method

Subjects

32 Cardiff University students (Mean age = 20, S.D. = 5.39) participated for either course credit or a payment of £5. None of the participants had taken part in any of the previous experiments and were given full credit/payment when the trials were completed in an attempt to maintain levels of effort across trials.

Design

There were two between subject factors of interface type and problem order. Like Experiment 2, interface type had two levels that consisted of a 1-Move control group and a Look-ahead group that allowed for the implementation of any number of legal moves at once.

Problem order consisted of 4 problems in total (see Appendix A), two problems with short solution paths (SH1 & SH2; 12 moves each in a set order) and two problems with longer solution paths (LG1 & LG2; 17 moves each in a set order). The short solution paths were given first to half of the participants followed then by the two longer problem solutions (SH Vs. LG). For the other half of participants this presentation order was reversed (LG Vs. SH). Participants were unaware of the differences in problems. The within subject factor of trial had 4 levels.

As before, total time to solution, inter-move latency times, number of moves, look-ahead depth measures, ratio of palindromic moves and moves minus palindromes were all taken as dependent measures. Additionally, number of trials solved was also included to examine successful completion of problems across trials, between interfaces users and by problem length.

Materials

The method of control for both interface groups was identical to that used in experiment 2. The scoreboard was located directly to the right of the 8-puzzle on the interface. The score would update in the same manner as in experiment 2 for the respective interface users. The goal state was once again present at all times on the screen along with the 8-puzzle. There were three key differences however in the materials used in experiment 3.

Firstly, before users attempted the 8-puzzle they were given additional instructions to the general instructions used to explain interface controls and the particular aim of the puzzle. These additional instructions constituted a hint and were given to all participants. The hint was designed with the intention of highlighting important features of the puzzle in the hope that it would increase the levels of conceptual knowledge that some participants appeared to lack, as demonstrated by performance in Experiment 2. This involved a worked example using sequences of pictures to highlight the solution to arrange tiles in a numerical order as opposed to concentrating upon simply organising them by absolute position. A copy of the full instructions can be found in appendix B. The aim of the hint which was given to all participants was to

try and alleviate the seemingly self imposed constraints that a subsection of participants place upon their problem solving behaviour.

The second change was that instead of completing the same problem over numerous trials, participants instead completed 4 different 8-puzzles, copies of which can be found in Appendix A. A set of two problems each with 17 move solution paths and a second set consisting of another two problems, requiring 12 moves to solution.

The third change was the implementation of a 10-minute time limit per problem. When 8 minutes had passed a 2-minute warning appeared underneath the 8-puzzle for 5 seconds warning them of the approaching time limit.

Procedure

The procedure was identical to the previous experiment with two exceptions. The aim of the 8-puzzle and the respective method of controls were explained to all participants as in experiment 2. When they had indicated they understood the aim of the experiment they were then shown a second screen containing the hint instruction (see Appendix B). Participants were told that when people attempt to solve the 8-puzzle there are a number of common mistakes that people often make. The first of these being that they sometimes refuse to move tiles that they feel are in their correct place and do not need to be moved again. Participants were told that in some cases such a presumption was incorrect and could lead to a problem never being completed. The second instruction aimed to increase performance was taken from the sequence of pictures presented on screen. They were shown one of the typical states visited by participants who were unable to solve the 8-puzzle. Again the need to move tiles was explained and a simple rotation mechanism could resolve such an issue. They were allowed to study the hint diagram sequence until they felt they understood the

concept. The 8-puzzle was then loaded and participants were told that they could press the 'Begin' button whenever they were ready. They were informed that they were to complete two simple rotation problems to begin with and that once these had been completed they would then begin the main stage of the experiment by solving four different 8-puzzles trials. All participants were informed that there was a 10 minute time limit per 8-puzzle trial and that they would receive a 2 minute warning on the screen when the 10 minute time limit was approaching. If a trial was not solved before the time limit expired participants simply clicked the 'Begin' button to start the next trial until all 4 trials had been attempted.

Results

The dependent measures were log transformed to stabilise for variance. Due to a number of trials not completed missing trial values were replaced with a grand trial mean calculated from both interface groups' participants who had finished the trial. All main effects were analysed using a 3-way mixed ANOVA unless stated.

Completion rates

From the 64 trials completed by each interface group a Chi Square test on total numbers of trials successfully completed revealed no significant difference between groups (Look-ahead 92.18% Vs 1-Move 93.84%).

Combining Interface Groups' success rates and analysing them by problem set (SH Vs. LG) revealed no difference in the success of problems solved based upon solution length. The short solution paths revealed a slightly higher completion rate of 96.87%,

while the long solution set problems had a combined solution rate of 90.62%, a Chi Square test again revealed no significant difference.

Total Moves

The number of moves made over trials by interface group were analysed and can be seen below in Figure 21.

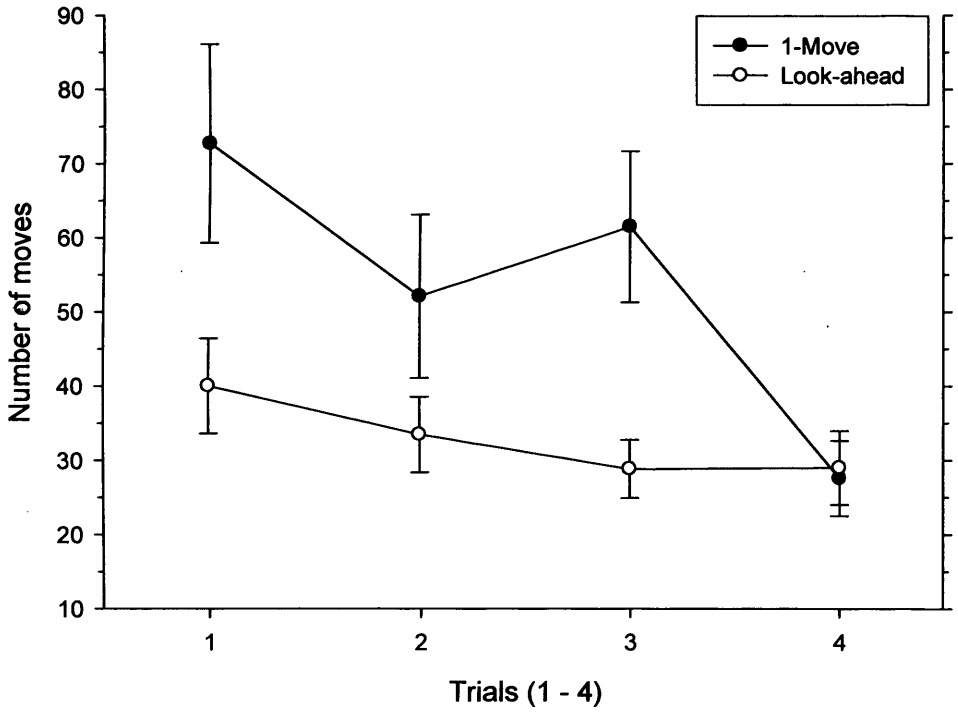


Figure 21. Total Moves to Solution for Interface Groups Across Trials

N.B. Each trial averages over two different problems

There was a significant effect of interface on number of moves made, $F(1, 28) = 5.196$, $p < .05$, $MSE = .75$, with those being encouraged to look-ahead completing trials in fewer moves than the 1-Move controls.

The main analysis also revealed a significant effect of trial, $F(3, 84) = 4.57$, $p < .005$, MSE = .35. The analysis also revealed a significant Trial x Problem Order interaction, $F(3, 84) = 4.87$, $p < .004$, MSE = .38, with problems in the short solution set being solved in fewer moves when shown after completing the longer solution path problems. Simple main effects analysis revealed the second problem in the short solution set being solved in fewer moves when it presented in the second set of problems compared to when solved in the first block of problems, $F(1, 28) = 14.03$, $p < .001$, MSE = .72. This suggests that performance on SH2 was at ceiling levels caused presumably by, not only the shorter solution path, but unlike SH1 also contained a more immediately visible solution path. Examining the solutions of trial 4 participants revealed that for the 1-Move condition out of the 8 participants who would have received the shorter solution path, 5 solved it in the minimum number of moves. There were also 3 out of 8 participants in the LG2 condition who solved it in the minimum number of moves. Similarly for the look-ahead condition, 4 participants solved the short solution path in the minimum number of moves while 2 participants solved the LG2 solution in the minimum number of moves. The remaining trials for the short solution path were solved in very few moves over the minimum leading to the observed high trial 4 performance.

Linear contrast revealed no significant linear contrast for the look-ahead group, $F(1, 14) = 2.03$, $p > .1$, MSE = .15. There was also no evidence of a quadratic component ($F < 1$). For the 1-Move group there was a significant linear, $F(1, 14) = 7.03$, $p < .02$, MSE = .83, but no quadratic component, $F(1, 14) = 1.69$, $p > .2$, MSE = .10.

Total Moves minus Palindromes

Palindromic sequences were once again calculated and subtracted from the number of moves made. The results are presented below in Figure 22 below. Similar to the previous experiment the results of removing palindromes from the total number of moves appears to have little overall effect on the analysis, again suggesting that interface groups are making similar amounts of backtracking moves.

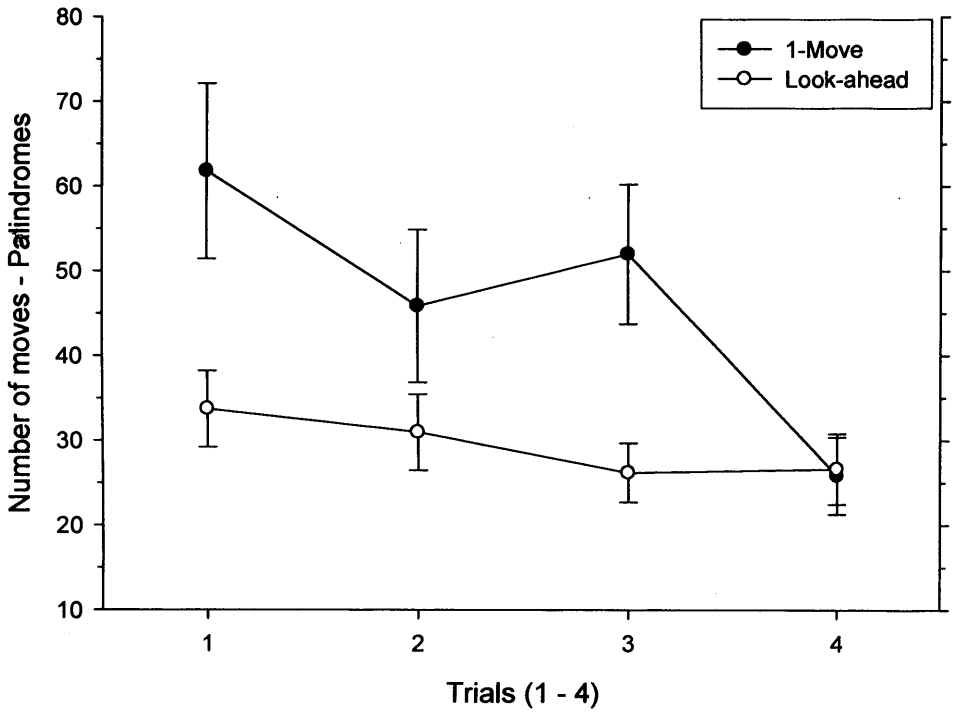


Figure 22. Total Number of Moves Minus Palindrome Moves

The analysis revealed a significant effect of interface on number of moves made, $F(1, 28) = 5.416$, $p < .03$, $MSE = .65$. There was also a significant effect of trial, $F(3, 84) = 4.03$, $p < .01$, $MSE = .27$, and as discussed above a significant Trial x Problem Order interaction, $F(3, 84) = 5.34$, $p < .002$, $MSE = .36$, with the second problem in the short

solution set being solved in a fewer moves when it was the last trial rather than when in the first set, $F(1, 28) = 15.46$, $p < .001$, $MSE = .66$.

Linear contrast analysis for look-ahead interface users revealed no linear trend, $F(1, 14) = 1.81$, $p > .2$, $MSE = .10$, and no quadratic component ($F < 1$). The 1-Move analysis revealed a linear trend, $F(1, 14) = 6.93$, $p < .02$, $MSE = .72$, and no quadratic component, $F(1, 14) = 1.35$, $p > .2$, $MSE = .07$.

Palindrome Ratio

The ratio of palindromic moves to total number of moves was calculated per trial and can be seen in Figure 23. As found previously, there was no effect of interface group on the ratio of palindromic moves in a participants solution attempt on any problem, $F(1, 28) = 2.19$, $p > .1$, $MSE = .03$.

The main analysis revealed a significant effect of trial, $F(3, 84) = 2.96$, $p < .05$, $MSE = .03$, but did not find any evidence of any Trial x Problem Order interactions or Trial x Interface interactions.

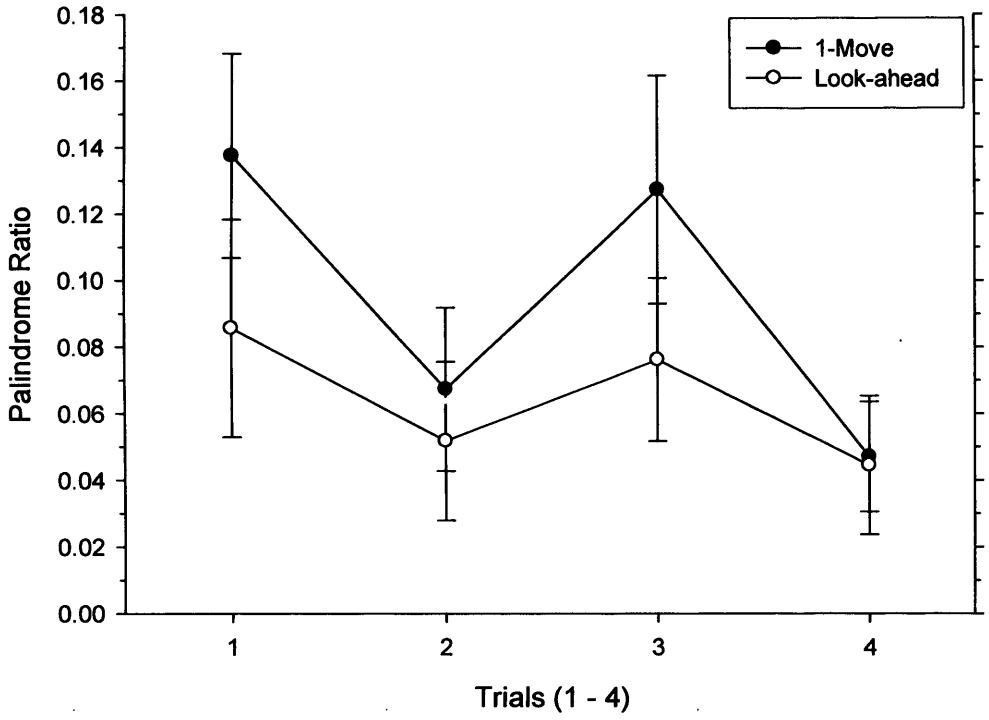


Figure 23. Ratio of Palindromes Contained in Solutions Across Trials

There were no linear or quadratic components for the look-ahead group (F's < 1). There was slight evidence of a linear trend for the 1-Move group, $F(1, 14) = 3.22$, $p > .09$, MSE = .04, but none for a quadratic component ($F < 1$).

Total Time

Examining total time to solution found no significant effect of interface on the time taken by participants to solve trials, $F(1, 28) = 3.08$, $p < .09$, MSE = 1.084.

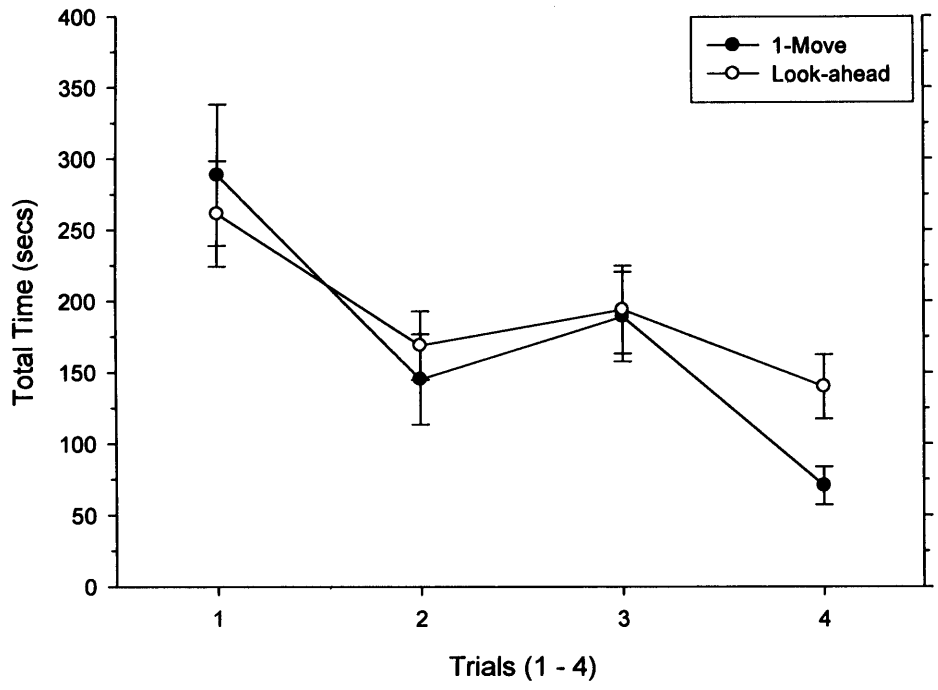


Figure 24. Total Time to Complete Problems over Trials

The main analysis revealed a significant effect of trial, $F(1, 28) = 11.78$, $p < .001$, MSE = 1.08. There was also a significant trial x order interaction, $F(3, 84) = 3.93$, $p < .02$, MSE = .36, with the second problem in the short solution set being solved quicker when it was the last trial rather than when in the first set and presented first, $F(1, 28) = 8.27$, $p < .008$, MSE = .70. No other interactions were significant.

Trend analysis revealed a linear component to the look-ahead groups performance, $F(1, 14) = 6.86$, $p < .02$, MSE = .54, but no quadratic component ($F < 1$). There was a significant linear component for the 1-Move group, $F(1, 14) = 15.51$, $p < .001$, MSE = 1.99, and no quadratic component ($F < 1$).

Inter-move Latency

Consistent with previous findings look-ahead interface users spent a much greater time planning their moves compared to those in the 1-Move control condition (Figure 25 below), $F(1, 28) = 30.00$, $p < .001$, $MSE = 2.70$.

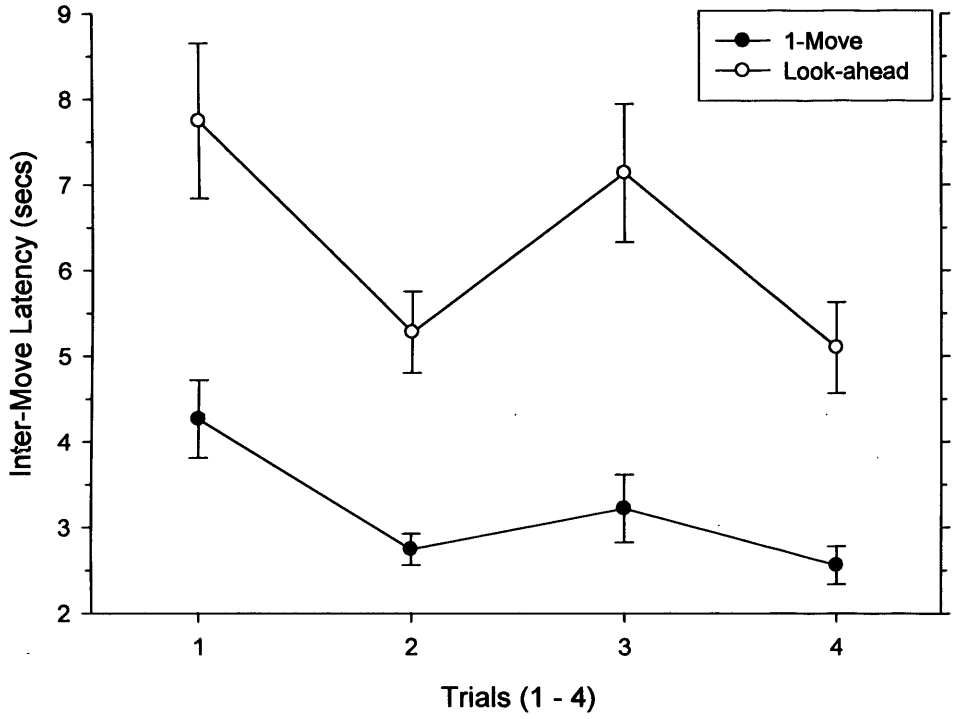


Figure 25. Inter-Move Latency Times over Trials for Both Interface Groups

There was a significant effect of trial, $F(3, 84) = 24.51$, $p < .001$, $MSE = .25$. The analysis revealed no evidence of any interactions.

A trend analysis on look-ahead performance revealed a significant linear component to the curve $F(1, 14) = 24.20$, $p < .001$, $MSE = .14$) but no quadratic component ($F < 1$).

For the 1-Move group there was a significant linear component to the curve, $F(1, 14) = 46.51$, $p < .001$, $MSE = .28$, and also a significant quadratic component to the curve ($F(1, 14) = 5.83$, $p < .05$, $MSE = .03$.

Look-ahead Measurement

The average look-ahead of subjects can be seen below in Figure 26. A repeated measures ANOVA on the look-ahead measure over trials revealed a significant effect, $F(3, 42) = 6.03$, $p < .002$, $MSE = .05$.

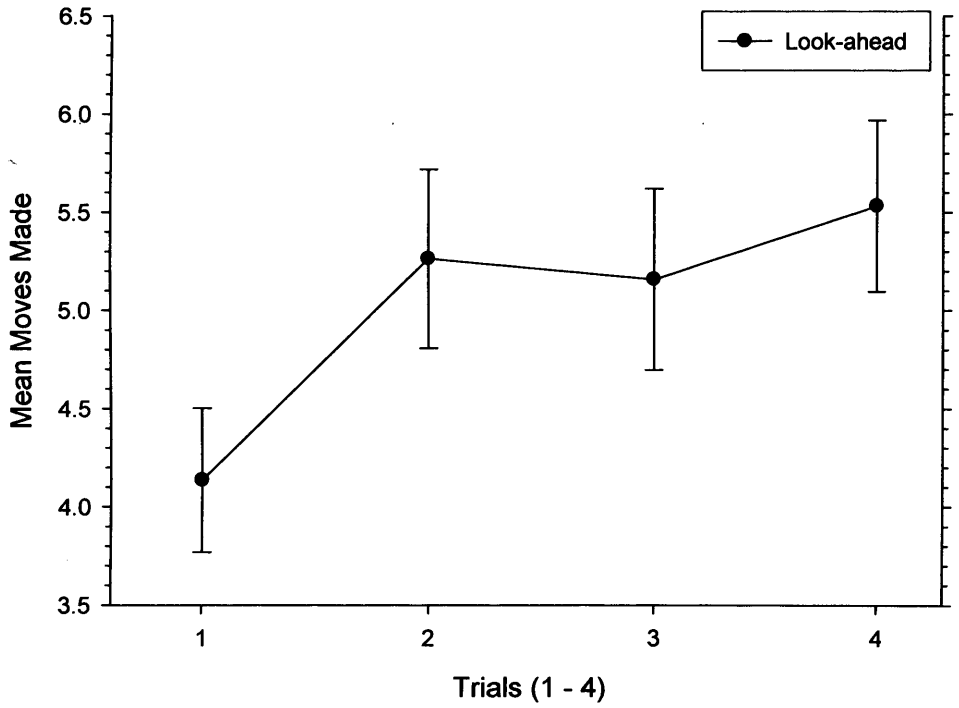


Figure 26. Number of Tiles Entered on Average over Trials by Look-ahead Interface Users

There was a significant linear component, $F(1, 14) = 7.56$, $p < .02$, $MSE = .12$, but no quadratic component, $F(1, 14) = 3.21$, $p < .095$, $MSE = .02$.

Discussion

The current experiment extended the previous findings by confirming previous results regarding both the benefits of look-ahead for performance and the gradual increase in look-ahead over trials. The combination of a hint intended to increase initial conceptual knowledge of the problem and a time limit have also further clarified the results by showing that the interface manipulation can work with problems of varying solution path lengths. Although the results are made somewhat more difficult to interpret due to order effects found for short solution path problems and possible ceiling effects, the results are on the whole positive.

Total time was once again unaffected by the increased look-ahead that look-ahead interface users were partaking in, with the increased benefits in performance leading to fewer moves needed making up for the increase in planning times. The lack of difference in the proportion of palindromic moves once again strengthens the assumption that the current look-ahead interface does not lead users to reduce backtracking moves perhaps due to the ease of undoing a move(s).

The numbers of trials completed were also high on average with the time limit providing an exit for participants that were unable to find a solution for the problems. Rather than being unable to complete all trials, the effects seem to be isolated cases, yet when they occur they can lead to long periods where a solution may be being searched for but is unable to be found.

Although initial look-ahead was much more elevated on trial 1 than in the previous experiment this is probably due to the increased performance on short solution path problems, with a number of participants immediately perceiving a possible solution path from the beginning of problem presentation and simply entering greater numbers of moves than would happen naturally.

A final experiment aimed to clarify and confirm the current results by extending the number of trials that participants are required to complete and by also using 8-puzzle problems that each take the same number of moves to complete. Presenting these trials in a completely randomized order will also answer questions regarding the transferability of skills and general learning benefits of increased look-ahead.

Experiment 4

The final experiment involving the 8-puzzle was carried out to confirm previous findings of the benefits of look-ahead, the use of a hint aimed at increasing conceptual knowledge for participants and to provide further measurement of the extent of look-ahead over a larger number of trials but with problems all requiring the same number of moves to be completed. The previous experiment used two different problem sets of differing solution lengths containing two problems each. Performance measures therefore were in some cases a little difficult to interpret due to possible order and evidence of ceiling effects for problems with short solution paths. These may have contributed to inflated levels of look-ahead as solutions may have been more immediately obvious and therefore have lead to large measures of look-ahead being recorded.

Method

Participants

24 Cardiff University students (Mean Age = 21.25, S.D. = 4.38) were paid £6 or given course credit for taking part in the experiment. None of the participants had taken part in any of the previous 8-puzzle experiments and were all given either full payment or credit upon completion of the trials.

Design

The current experiment had interface type as the only between subject factor with two levels in which participants completed the 8-puzzle using either a 1-Move or Look-ahead Interface. The within subject factor of trial had 8 levels consisting of 8

different 8-puzzles, each requiring 17 moves to solution. As in Experiment 3, a hint aimed at increasing successful completions and 10-minute time limit per trial was also imposed for all participants.

As in previous experiments the total number of moves to solution, total time, latency time, the ratio of palindromic moves, number of moves minus palindromes and number of trials successfully completed were recorded for both groups. A look-ahead measure determined by number of tiles entered on average was also recorded for Look-ahead interface users.

Materials

The materials used were identical to the previous experiment in terms of hint presentation, interface manipulation and a 10-minute time limit per trial. When two minutes remained on a trial, a warning appeared on screen telling participants of the approaching time limit.

The 8 different problem start states requiring 17-moves each to solution were generated using a PROLOG program developed by Bratko (2000) and can be found in Appendix C. A simple rotation problem was also included requiring only 5 moves for participants to complete before they began solving the experimental trials. All 8-puzzle problem states were presented in a randomized order generated by the computer program. Participants were unaware of the lengths of any of the solutions.

Procedure

Participants were introduced to the 8-puzzle and told of the aim to transform the given start state into the goal state a picture of which was placed on screen. When participants indicated they fully understood the aim of the task they were then given instruction about their respective method of controls. The hint screen was then presented and participants informed of some of the typical constraints that people appear to impose upon themselves and a simple method of removing obstacles if they should find themselves in such a position. The 8-puzzle screen was then loaded and participants were told to press the 'Begin' button upon which the trial would start. They were told that they would be solving 8 different problems and that there was a 10 minute time limit per problem. If they failed to solve any problem within the allotted time they would be informed by a message on the screen upon which they would simply click the 'Begin' button again to begin the next trial until all 8 problems had been attempted.

Results

Trials with missing values due to not being completed were replaced with grand trial means calculated over both interface groups from participants' data who had successfully completed the trial.

All data in the main analysis were analysed using a 2-way mixed ANOVA with interface as the between subject factor and trial as the within subject factor. Dependent measures were log transformed to stabilise for variance.

Pass/Fail Rates

From the 96 trials attempted by all participants in each interface group the number of successful completions was recorded and can be seen below in table 2.

Table 2

Pass/Fail Rates for Participants solving 8 Problems

	1-Move	Look-ahead
Pass	81	85
Fail	15	11
Total	96	96
Pass %	84.375%	88.54%

A Chi Square test revealed no differences between interface groups on the number of trials successfully completed. The completion rate data was collapsed for the two interface groups and number of trials successfully completed over the 8 trials were broken down into 2 sets of 4 (1^{st} 4 trials Vs. 2^{nd} 4 trials) to compare possible differences in numbers of trials completed with increasing experience of the problem. Chi square revealed a significant difference ($Chi = 4.45$, $df = 1$, $p < .05$), with more trials being successfully completed on the second block of trials.

These results are comparable with the previous experiment, with high completion rates being observed for both groups overall.

Total Moves

The number of moves to solution per trial was analysed (see Figure 27 below), revealing a significant effect of Interface on total number of moves, $F(1, 22) = 15.11$, $p < .001$, $MSE = 1.52$.

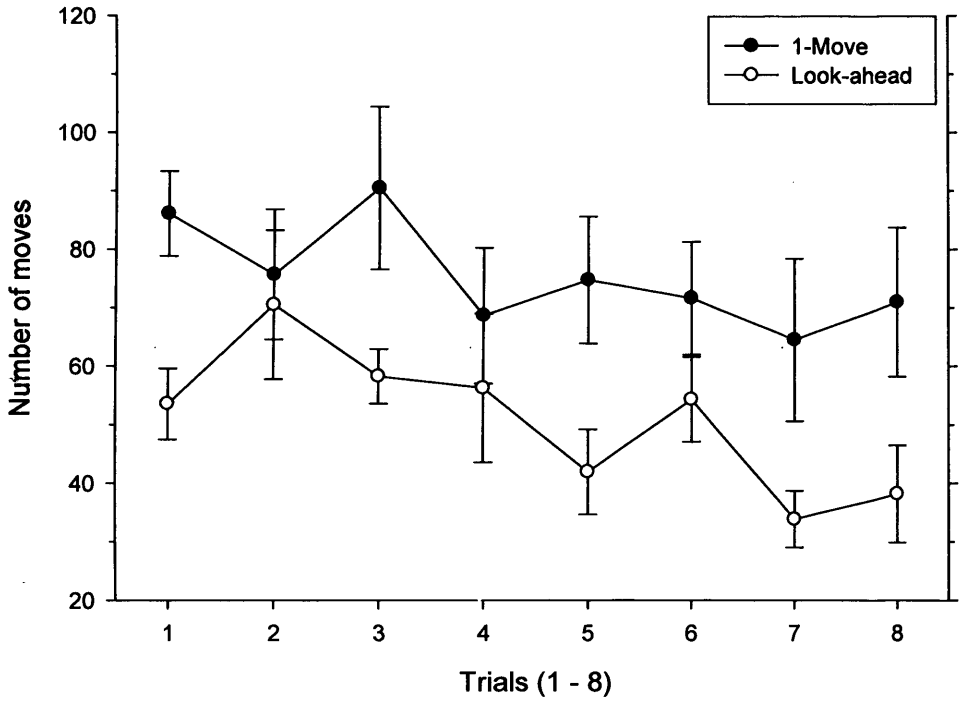


Figure 27. Total Number of Moves to Solution over Trials

There was also a significant effect of trial, $F(7, 154) = 2.846$, $p < .01$, $MSE = .16$. The analysis did not reveal any trial by interface interaction ($F < 1$).

A trend analysis over trials revealed a significant linear component to the curve for those in the look-ahead condition, $F(1, 11) = 9.43$, $p < .02$, $MSE = .65$. There was no significant quadratic trend to the curve, $F(1, 11) = 2.70$, $p > .1$, $MSE = .05$.

Contrastingly, a trend analysis on 1-Move performance revealed neither a linear component, $F(1, 11) = 2.36$, $p > .1$, $MSE = .21$, nor a quadratic component ($F < 1$) to the curve.

Total Moves minus Palindromes

Palindromic sequences were analysed from moves made during a trial and subtracted from the number of moves, the results are shown below in Figure 28.

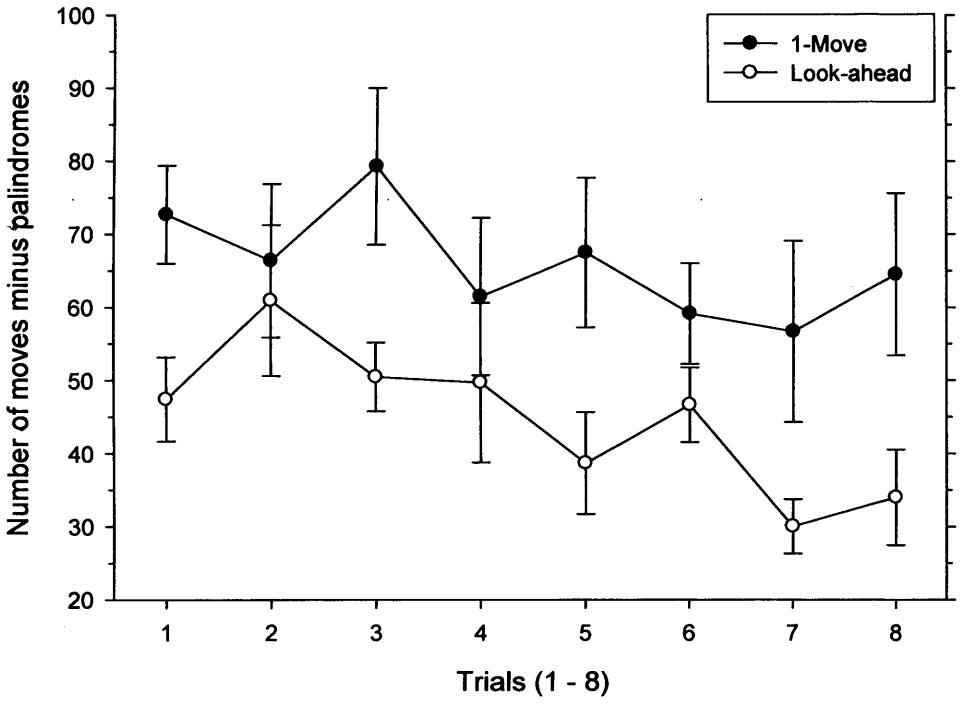


Figure 28. Total Number of Moves Minus Palindromes Over Trials

Similar to the above analysis, there was a significant effect of interface, $F(1, 22) = 13.25$, $p < .001$, $MSE = 1.42$. There was also a significant effect of trial, $F(7, 154) = 2.48$, $p < .02$, $MSE = .13$, and once again no trial by interface interaction ($F < 1$).

There was a significant linear component to the curve for the look-ahead group, $F(1, 11) = 9.70$, $p < .01$, MSE = .52. and no quadratic component, $F(1, 11) = 1.59$, $p > .1$, MSE = .04. For the 1-Move group there was no evidence of a linear component to the curve, $F(1, 11) = 1.89$, $p > .1$, MSE = .16, or a quadratic component ($F < 1$).

Palindrome Ratio

Total proportion of palindrome moves involved in move sequences was also calculated and can be seen below in Figure 29.

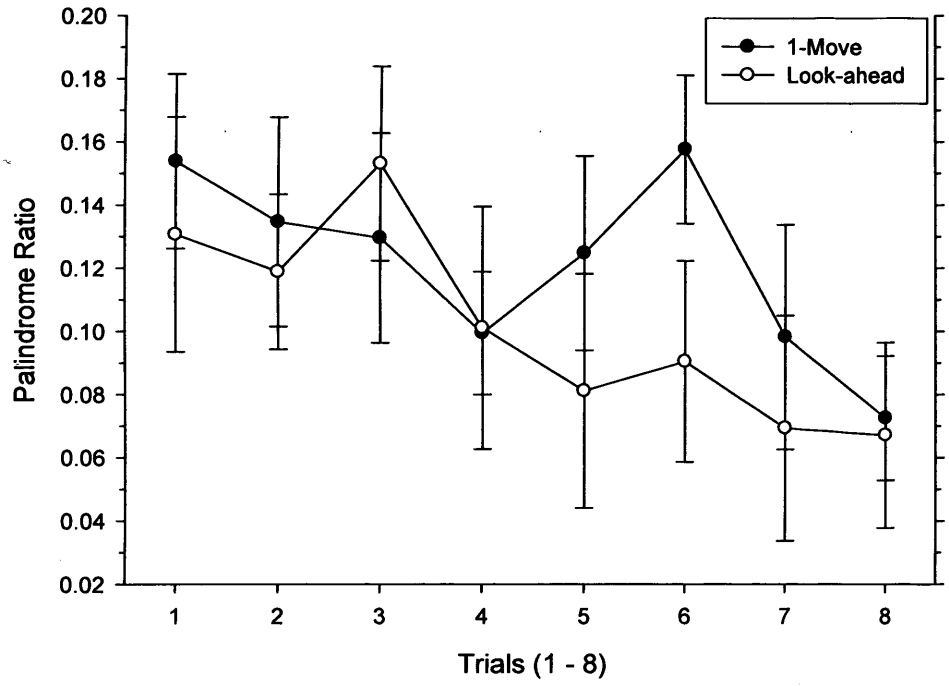


Figure 29. Ratio of Palindromes to Total Moves Across Trials

Analysis revealed no effect of interface on the ratio of palindromes moves made, $F(1, 22) = 1.30$, $p > .1$, MSE = .02. There was also no effect of trial, $F(7, 154) = 1.51$, $p > .1$, MSE = .02, and no evidence of a trial by interface interaction ($F < 1$).

Trend analysis revealed no significant linear trend, $F(1, 11) = 3.68$, $p < .1$, $MSE = .06$, and no quadratic component for the look-ahead condition ($F < 1$). For the 1-Move group there was also no linear component although there was some evidence of a slight trend, $F(1, 11) = 3.58$, $p < .09$, $MSE = .03$. There was no indication of a quadratic component ($F < 1$).

Total Time

The total time to complete trials was analysed and as in previous findings the ANOVA revealed no effect of interface on time taken to complete trials ($F < 1$).

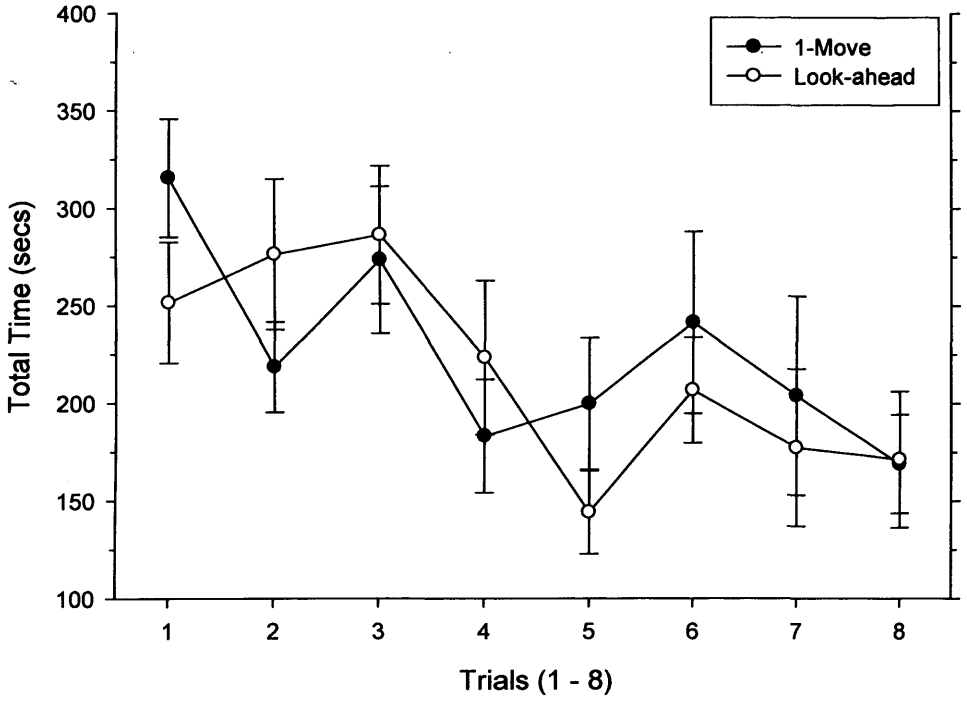


Figure 30. Total Time to Solution Across Trials for Interface Groups

As shown in Figure 30, there was a significant effect of trial, $F(7, 154) = 4.74$, $p < .001$, MSE = .32, with participants completing trials faster by the end of the session. There was again no evidence of any interaction between interface and trial ($F < 1$).

Trend analysis for those in the look-ahead group revealed a significant linear component to the curve, $F(1, 11) = 11.80$, $p < .01$, MSE = .99, but no quadratic component ($F < 1$). For the 1-Move group there was also a significant linear component to the curve, $F(1, 11) = 4.953$, $p < .05$, MSE = .71, and like the look-ahead condition no indication of a quadratic component to the curve, $F(1, 11) = 1.02$, $p > .1$, MSE = .05.

Inter-Move Latency

As Figure 31 clearly highlights there was a very significant effect of interface on time taken to make a move. The analysis on inter-move latency times revealed a significant effect of interface, $F(1, 22) = 10.31$, $p < .004$, MSE = 1.43) on time taken to enter moves. There was also a significant effect of trial, $F(7, 154) = 4.28$, $p < .001$, MSE = .07. There was no interaction between trial and interface ($F < 1$).

A trend analysis revealed no linear, $F(1, 11) = 1.26$, $p > .1$, MSE = .06, or quadratic components to the curve, $F(1, 11) = 3.11$, $p > .1$, MSE = .06, for look-ahead interface users.

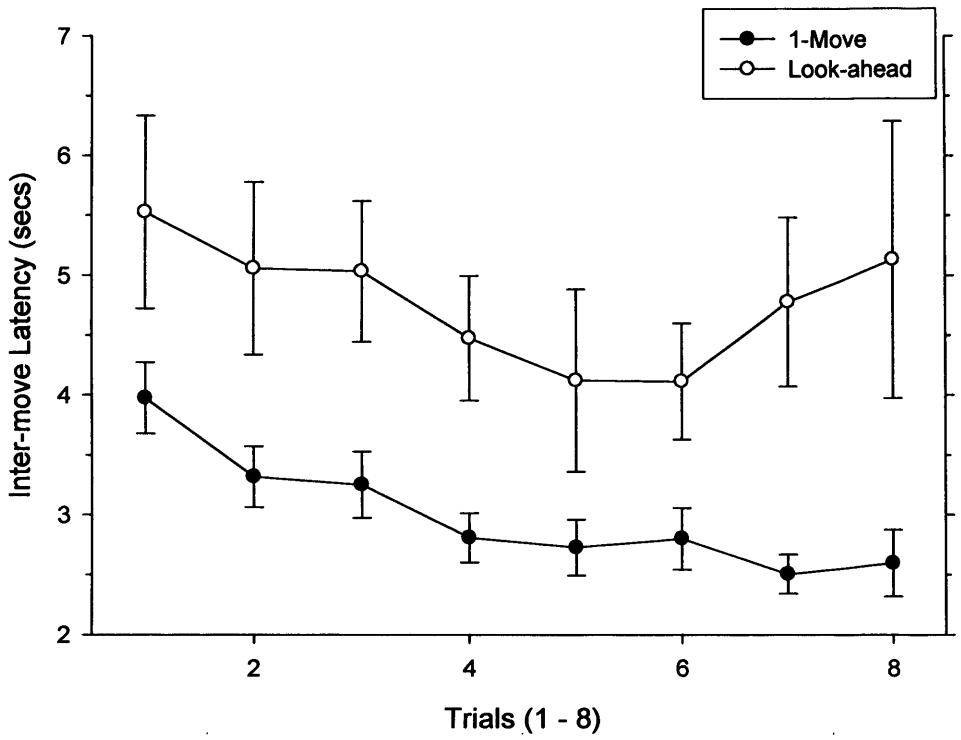


Figure 31. Inter-Move Latency Times Over Trials for Both Interface Groups

For the 1-Move interface users there was a significant linear trend to the curve, $F(1, 11) = 20.43$, $p < .001$, $MSE = .33$, and also a significant quadratic component, $F(1, 11) = 16.40$, $p < .002$, $MSE = .03$.

Look-ahead Measurement

Mean number of tiles entered per go was recorded per participant and can be seen below in Figure 32. A repeated measures ANOVA over trials revealed a significant effect of trial, $F(7, 77) = 5.51$, $p < .001$, $MSE = .08$.

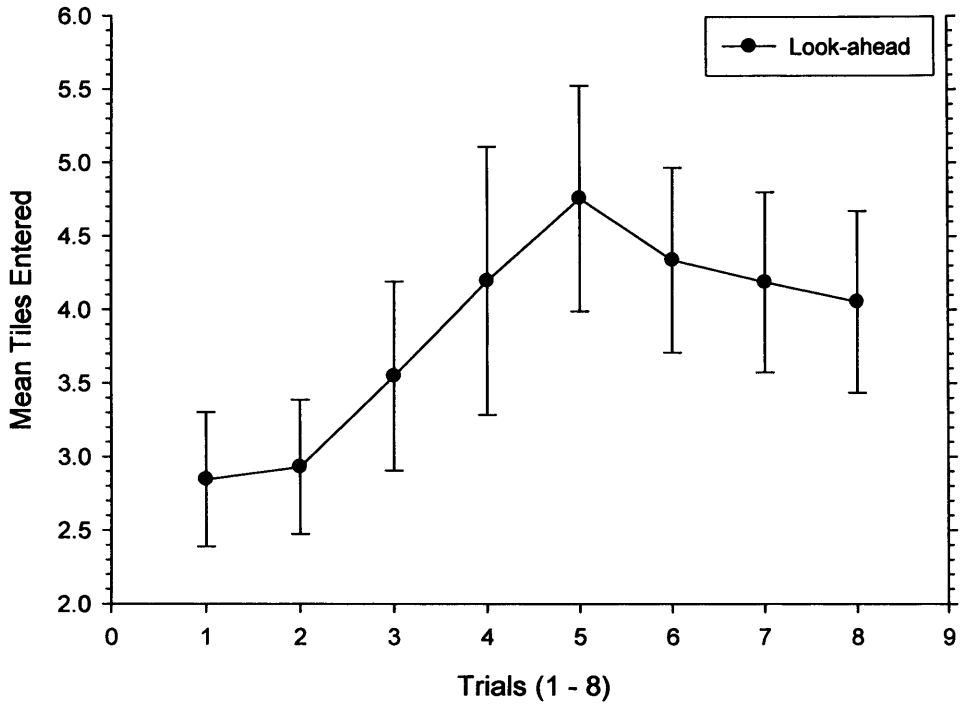


Figure 32. Mean Number of Tiles Entered Per Move for Look-ahead Interface Users

As previous results have demonstrated there was also a significant linear component, $F(1, 11) = 7.55$, $p < .02$, $MSE = .38$, and a significant quadratic component, $F(1, 11) = 6.20$, $p < .05$, $MSE = .11$, to the curve.

Discussion

The results once again have confirmed a number of previous findings. The manipulation to increase look-ahead once again lead to better problem solving performance over trials. This importantly did not come at the price of significantly increasing the amount of time spent problem solving. The look-ahead manipulation

lead to longer latency times which are typically indicative of increased levels of planning. Palindromic moves were again unaffected by the manipulation although there was a slight tendency for the number of backtracking moves to diminish over trials although this was not significant.

Look-ahead also increased over trials as expected and the evidence seems to suggest that there is a natural plateau of approximately 4-moves that participants will look-ahead while solving novel levels, after a sustained period of time and practice.

The lack of a difference in number of trials solved also suggests that the planning manipulation is applicable to all levels of planners due to its inherent flexibility, with the onus on greater planning when participants themselves see the opportunity rather than the strict enforcing of look-ahead which may exclude portions of the population. The trial 1 look-ahead values of 2.8 tiles per go again lend support to the conclusions from experiment 1 regarding the difficulty of simulating 3-moves at a time when a task has just begun. Immediately asking all participants to move 3 tiles at a time was a likely factor preventing a large proportion of the population from operating at their natural planning limits within the novel problem solving environment and that lead to such poor performance on trial 1 of Experiment 1 particular.

Chapter 3

Look-ahead Manipulations and Water Jar Problems

Introduction

The argument that greater look-ahead will have positive outcomes for performance is largely supported by the performance and look-ahead span data reported in the previous chapter. To further qualify these assumptions and investigate their possible application to a new problem solving domain would advance this presumption further and increase knowledge of the look-ahead component of problem solving. Transferring the previous successful manipulation to an unrelated problem, differing in its representational characteristics and task demands, would demonstrate its versatility and general applicability as a worthwhile performance aid. The new class of problems that matched the necessary criteria are widely referred to as 'Water Jars' problems.

Experiment 5

One of the first experiments to report on Water Jars problems was conducted by Luchins (1942), who trained one group of subjects on a number of water jars problems using a specific solution path to establish a particular automated mental "Einstellung" or as it is commonly referred to "set". When faced with new problems, the overly complex paths continued to be used even when more efficient solution paths were available. In contrast, a control group were able to identify these more direct solution paths as they were uninfluenced by a previous mental set.

Previous models of water jars performance were divided in their assumptions regarding the likely depth of look-ahead search during problem performance. Atwood & Polson (1976) argued that due to fundamental working memory limitations in terms of cognitive capacity and processing, a large look-ahead span would be unlikely. To support this argument they developed a model of water jars performance constrained by a look-ahead depth of one step that could still accurately model typical human performance. Yet more recently and as previously described, participants in a study by Delaney et al. (2004) were instructed to plan the complete solution path before being allowed to implement any moves. A filler task to prevent active plan rehearsal, placed between when participants indicated they were ready to implement their solution and the time they actually began entering moves, had little effect upon number of moves to solution. This result was taken as evidence that memory for a solution path was not solely responsible for the improvement in performance when planners were compared with control participants. No specific interface manipulations were at play, but rather a simple verbal instruction to refrain from solving the problem until fully satisfied the correct solution path had been successfully pre-planned.

There is however a possible methodological weakness in the design employed by Delaney et al. (2004). The planning instruction may have led to the solving of a puzzle any number of times before an indication that a plan had been formulated was given. Differences between experimental groups' performance may have simply been the result of the planning group having generated, regenerated and stepped through the solution path multiple times before indicating their ability to solve the problem. Although the filler task between pre-plan and implementation was designed to prevent any subsequent rehearsal, having the opportunity to simulate the solution path numerous times may have aided in bypassing or reducing the impact of the filler task.

Delaney et al. (2004) appear to have assumed that the solution will only have been simulated infrequently at best. If this is not the case, then the results may simply indicate therefore that participants are better at solving a water jars problem when it has been solved on more than one occasion. In contrast, the current manipulation aims to increase performance through the gradual progressive deepening of search. Any partial plans generated by look-ahead interface users should be more embedded in the first solution attempt rather than on an attempt simulated multiple times previously. If the verbal instruction used by Delaney et al. can be successful in improving performance, then it seems both logical and highly likely that an interface based manipulation will also be successful. This may be especially true if the manipulation supports a participants decision to total-order plan, or the interleaving and use of greater quality partial planning, when unable to form total-plans.

There are a number of possible predictions for performance data in the current experiment for total moves, total solution time and inter-move latency times. The predictions for number of moves to solution are two fold for Look-ahead interface users. Firstly, fewer numbers of excess moves should be required to reach the goal-state compared to 1-Move performance. Secondly, those using the Look-ahead interface should also solve a greater number of problems in the minimum number of moves. While Delaney et al. (2004) did not report optimal trial data, it would appear logical to assume that the intended increase in look-ahead and planning should lead to a larger number of trials being solved optimally.

Predictions for total solution time are more difficult to make. The complexity of generating one move in the Water Jars is vastly different from that needed to make one move in the 8-puzzle. The number of decisions required to make a single move is

much more complex, with the effect of implementing an operator on the next state not being immediately perceptually available, unlike the 8-puzzle. Information about the amount of water in the current jar and the contents of the jar that water is intended to be poured into must first be compared to check for legality. Furthermore, the subsequent effects on the contents of two sets of jars must also be calculated and finally how the current states of the three water jars stand in relation to the goal state also has to be considered when planning with the intention of performing with relative efficiency. This combinatorial complexity may only increase with the greater consideration of moves as the level of search deepens. Compared with the typically less planful performance of 1-Move users, less care may be taken if the inclination to plan or perform well is at a low enough level. The reduced likelihood of deeper search would also reduce the impact of the underlying complexity of move choice. Therefore, total solution time may be much greater in the current experiment for look-ahead users compared to controls. Total solution time data for the planning group was not reported by Delaney et al. (2004), making predictions for the current plan group also more difficult. The fact may still remain that total time to solution will remain unaffected by increased planning, with the improved performance once again compensating for the increased time spent planning. The effect of problem trial is however likely to play a role in the time taken to reach solutions across trials. It is not necessarily clear as to how total time will be affected by number of moves to solution, as the total number of moves in a solution path does not necessarily predict problem difficulty and therefore total time.

One consistent finding that will most likely remain unaffected by problem characteristics is in relation to inter-move latency times. In line with previous findings there should be strong evidence of an effect on inter-move latencies between groups,

with 1-Move controls having shorter latency times than their Look-ahead counterparts.

Look-ahead performance in terms of span is difficult to predict in comparison with the levels recorded from the 8-puzzle experiments. The need to combine several sources of information for just one move may mean that significant increases in look-ahead will simply not be possible, although the success of Delaney et al's planning manipulation suggests otherwise. Planners may adopt a more advanced problem solving strategy such as an algebraic method to cope with the increased load on working memory of increased search. Verbal protocols taken from planners by Delaney et al. (2004) found this a viable means by which to cope with the increased demands made by the planning requirement. Alternatively, the chunking of moves as evinced in earlier experiments may also alleviate the demands placed on working memory and allow for increased look-ahead when coupled with increased exposure to the task.

In contrast to 8-puzzle data, palindromes cannot be measured in the Water Jars problems, or at least very infrequently. Undoing a previous move may not actually return the problem to its previous state but create a totally new problem state. A more appropriate measure that can be recorded is number of *resets* made during attempted solutions. This is a move that return the contents of Jar A to its original contents and leaves Jars B and C empty (i.e. the original start states). A greater number of resets may be indicative of poorer planning (Knowles & Delaney, 2005), and may be likely when a decision is made that the current state is unlikely to result in a successful solution. Frequently restarting the puzzle to begin searching for a new, more promising solution path may be a mechanism by which to reduce the effort required to

generate plans. Fewer numbers of resets may therefore indicate greater planning, negating or reducing the need to return to the start state and begin again as the moves are being mentally constructed rather than in the physical space of the problem. A second reason to predict that fewer numbers of resets will occur for Look-ahead users is what could be considered a *natural cost* associated with making an incorrect move. A poor move choice may automatically be punished, as the nature of the puzzle ensures that allowing immediate backtracking to the previous state can rarely happen. Therefore, look-ahead users, when planning a greater number of moves, will most likely experience greater feelings of cost as depth of search and planning continues. The possibility of entering an undesirable state, which may require the entire process to begin again, may be enough to reduce the likelihood of resets. 1-Move users should be less inclined to feel such increasing costs and therefore a difference would be predicted.

Method

Participants

85 participants from Cardiff University participated for either course credit or a payment of £6. Participants were firstly screened for arithmetic ability as suggested by Delaney et al. (2004) by completing four mental arithmetic questions, a copy of which can be found in Appendix D. Two attempts per question were permitted and those answering two or more questions incorrectly, 37 in total, were released from the experiment. The remaining participants, 48 in total (M = 21.52 years, S.D. = 3.3), with 24 in each experimental group took part in the Water Jars phase of the experiment.

Design

The between subject variable of Interface consisting of 2 levels, had subjects using either a 1-Move control interface or a Look-ahead interface throughout all trials. A within subject variable of trial contained 8 different levels, details of which can be seen in table 3 of the materials section below. Dependent measures recorded for both interface groups included proportion of trials solved using the minimum solution path, number of excess moves above the minimum possible, number of resets to start state during trials, total time to solution and inter-move latencies. Additionally, for Look-ahead interface users the number of moves specified per implementation (i.e. Look-ahead span) and numbers of errors made during move implementation were also recorded by the program.

Materials

All materials, mental arithmetic screener and Water Jars program, were created using the Visual Basic 6.0 programming language and presented to participants on a 17" Sony CRT monitor.

The interfaces, both entirely mouse driven for all participants, were almost identical in both appearance and the method of control in terms of the processes involved in the selection and implementation of a move(s). The 1-Move interface used by control participants is shown below in Figure 33.

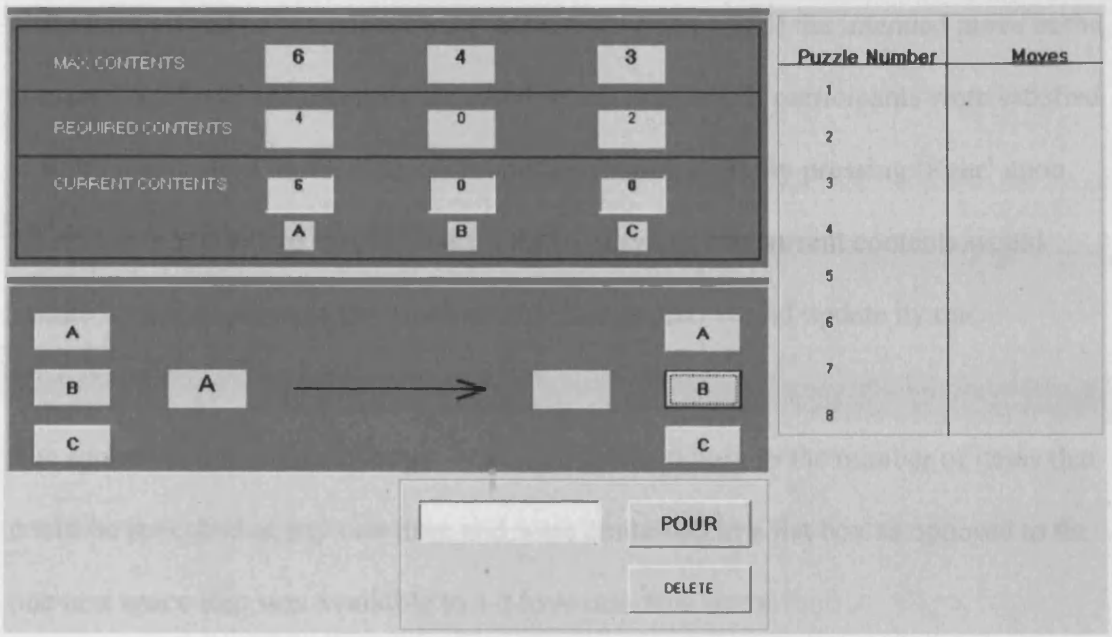


Figure 33. Screenshot of the 1-Move Interface Used in the Current Experiment

The max contents, required contents and current contents fields were left blank until the participant indicated they were ready to start a level by pressing a button marked 'Begin' that was situated just under the main puzzle. Upon this selection the required information (i.e. information about current state, start state and required goal state) for the current level appeared in the appropriate fields and the participant could begin implementing moves when they wished. The specification of a particular move followed a simple three stage process. Firstly, the jar from which water was to be poured must be specified. Participants selected from one of the buttons (A, B or C), situated to the left of the arrow, that indicated the jar water would be poured from. The appropriate caption would then appear on the left as shown by the caption 'A' in Figure 33 above. At any point before both caption boxes have been specified a participant may select another jar upon which the caption would automatically update. The second stage involved specifying the jar that water was to be poured into by again selecting either from buttons A, B or C, this time located to the right of the arrow.

Once both jars have been specified (i.e. both captions filled) the intended move in the form of 'A into B' for example appeared in the text box. If participants were satisfied with their specified move it could be implemented by simply pressing 'Pour' upon which the move would be checked for its legality and the current contents would update appropriately and the score for the current trial would update by one.

The look-ahead interface (Figure 34 below) differed only in the number of items that could be specified at any one time and were contained in a list box as opposed to the one text space that was available to 1-Move interface users.

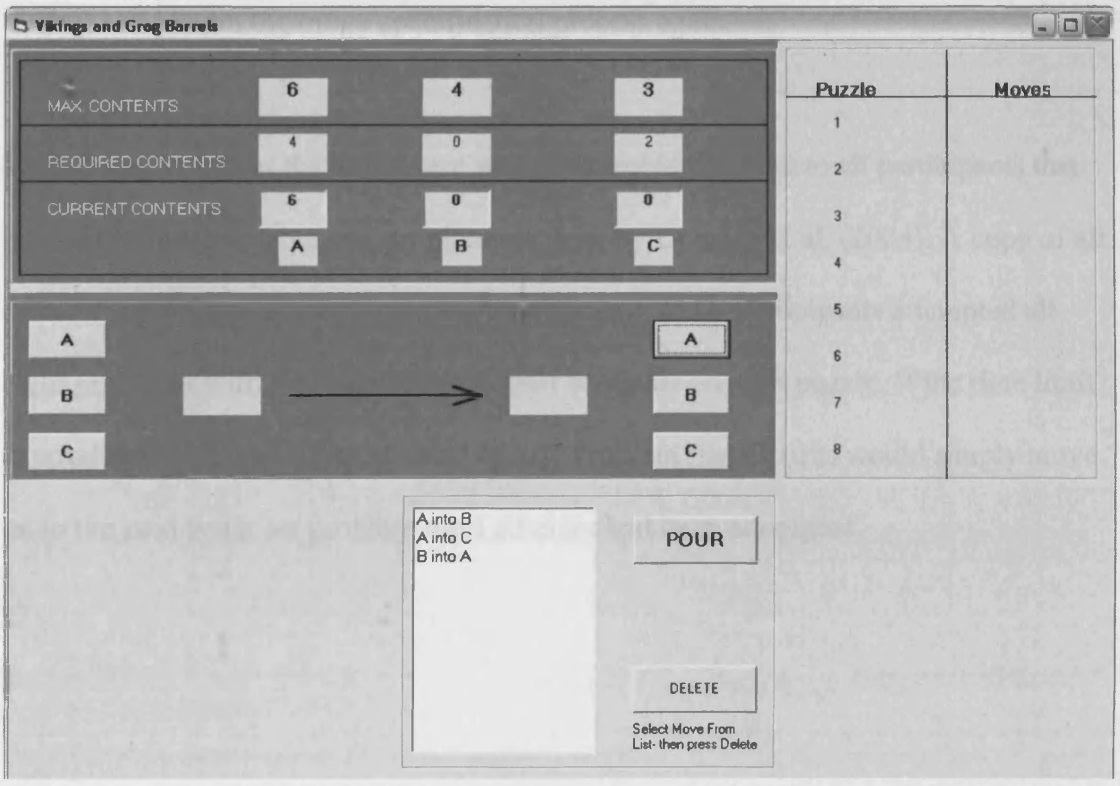


Figure 34. Screenshot of the Look-ahead Interface Used in the Current Experiment

A list of pre-planned moves specified at the users own discretion could be listed by those using the look-ahead interface and would be implemented in the order they appeared in the list when 'Pour' was pressed.

For all participants, all move(s) had to be logically possible for the current contents to change or for the scoreboard to update. Upon the detection of any moves that were not possible a message box would inform participants that a move could not be implemented and to check their move(s) again. The current contents would not update unless a move or sequence of moves were possible in their entirety. All participants could remove their intended choice from the text box by selecting it and pressing 'Delete' and begin the move specification process again.

Table 3 below shows the 8 different water jar problems given to all participants that have all been taken from test problems as used by Delaney et al. (2004). A copy of all Water Jar problems can be found in Appendices E to G. Participants attempted all eight problems with a 12 minute time limit being set on each puzzle. If the time limit expired before solution was reached on any problem participants would simply move on to the next water jar problem until all eight had been attempted.

Table 3

Details of the Water Jar Problems Used in the Current Experiment

Problem	Jug Sizes	Initial State	Goal State	Solution Length
1	6 / 4 / 3	6 / 0 / 0	4 / 0 / 2	3
2	5 / 4 / 3	5 / 0 / 0	4 / 1 / 0	3
3	9 / 7 / 2	9 / 0 / 0	5 / 4 / 0	4
4	8 / 5 / 3	8 / 0 / 0	2 / 5 / 1	4
5	11 / 6 / 5	11 / 0 / 0	4 / 6 / 1	5
6	10 / 5 / 4	10 / 0 / 0	4 / 5 / 1	5
7	10 / 7 / 3	10 / 0 / 0	9 / 0 / 1	6
8	11 / 8 / 3	11 / 0 / 0	2 / 8 / 1	6

Due to the problems being presented in a fixed order, interpreting results are somewhat more complicated. There is evidence however that randomizing the presentation order of problems may further complicate any interpretations, due to the bias introduced by solving simpler problems first versus participants who encounter more difficult problems first (Luchins & Luchins, 1990). Solving difficult problems first may negatively affect subsequent performance on simpler trials which may have added another dimension of difficulty making any analysis of the data almost not possible. The order was therefore was kept constant to avoid creating possible order effects of easy versus hard puzzles and for ease of comparison with previous work (cf. Delaney et al., 2004).

Procedure

Participants were first informed that they would be required to complete 4 mental arithmetic questions, presented one at a time on a computer screen. The importance of getting all questions correct was stressed to all participants. Participants would work out the answer to a question, enter their response into a text box and press the 'Return' key. If the answer was correct they were informed via a message box and would click 'OK' to move on to the next question until all 4 questions had been completed. If on the first attempt to answer a question the response was incorrect they were informed by an error 'beep' from the computer, their previous answer would be cleared from the text box and they would have a second opportunity to calculate the answer. If their second attempt was also incorrect they were informed that they would simply be moving on to the next question. When all 4 questions had been attempted participants with two or more incorrect responses were removed from the experiment.

Participants who were successful in completing the mental arithmetic phase were told they would be solving a number of problems using Water Jar puzzles. The concept of the puzzle was explained to participants and the main screen containing the interface was then loaded upon which further examples were given. When participants could clearly explain the aim of the puzzle back to the experimenter and indicated they clearly understood the mechanisms by which water could be measured and transferred the method of controls and scoreboard were explained to all participants. All participants were told to solve the puzzle to the best of their ability and the aim was to minimize the score per level. They were informed that all levels were different and that they differed in difficulty. Participants were also told of the time limit but that its purpose was to provide a means of exiting a problem should a solution not be

reached within a reasonable time limit. When participants were ready they started by pressing the 'Begin' button and finished when all 8 trials had been attempted.

Results

General Performance Data

Examining overall completion rates for both interface groups revealed a high success rate, with 90.88% of all trials successfully completed. From the 192 trials attempted by both experimental conditions, 1-Move interface users completed 91.6% of all trials whilst Look-ahead users completed 90.1% of trials. A Chi-square test revealed this difference not to be significant ($\chi^2 = .28$, *n.s.*, df = 1). Combining both interface groups' successful completion data revealed pass rates of 100%, 100%, 100%, 87.5%, 87.5%, 85.41%, 97.91% and 68.75% for trials 1 – 8 respectively.

As previously discussed, a measure of the number of errors during plan implementation was recorded for look-ahead users. The results showed that proposed plans were largely error free. The modal number of errors across trials and by participant was zero, indicating that the plans being generated were highly accurate and so will not be examined further.

Optimal Problem Performance

Examining 1-Move performance revealed that 65 trials from a possible 192 were solved in the minimum number of moves (33.85%). 89 trials were solved in the minimum number of moves by participants in the look-ahead condition (46.3%). A Chi-Square between interface groups on the total number of trials solved in the

minimum number compared to trials not solved in the minimum number of moves revealed this difference to be significant ($\chi^2 = 6.24$, $p < .03$, $df = 1$).

Number of Excess Moves

Excess moves, m, were $\log_{10}(m + 1)$ transformed to stabilise the variance for the purpose of analysis. Missing trial values were replaced with the grand mean $\log_{10}$ values (cf. Delaney et al., 2004). Total time and move latencies were also $\log_{10}$ transformed for the purposes of analysis with missing values being replaced in the same way as excess moves.

Number of excess moves, total time and inter-move latency were analysed using a two-way mixed ANOVA with Interface as a between subject factor and problem trial as the within subject factor.

The number of excess moves above the minimum possible per trial can be seen below in Figure 35.

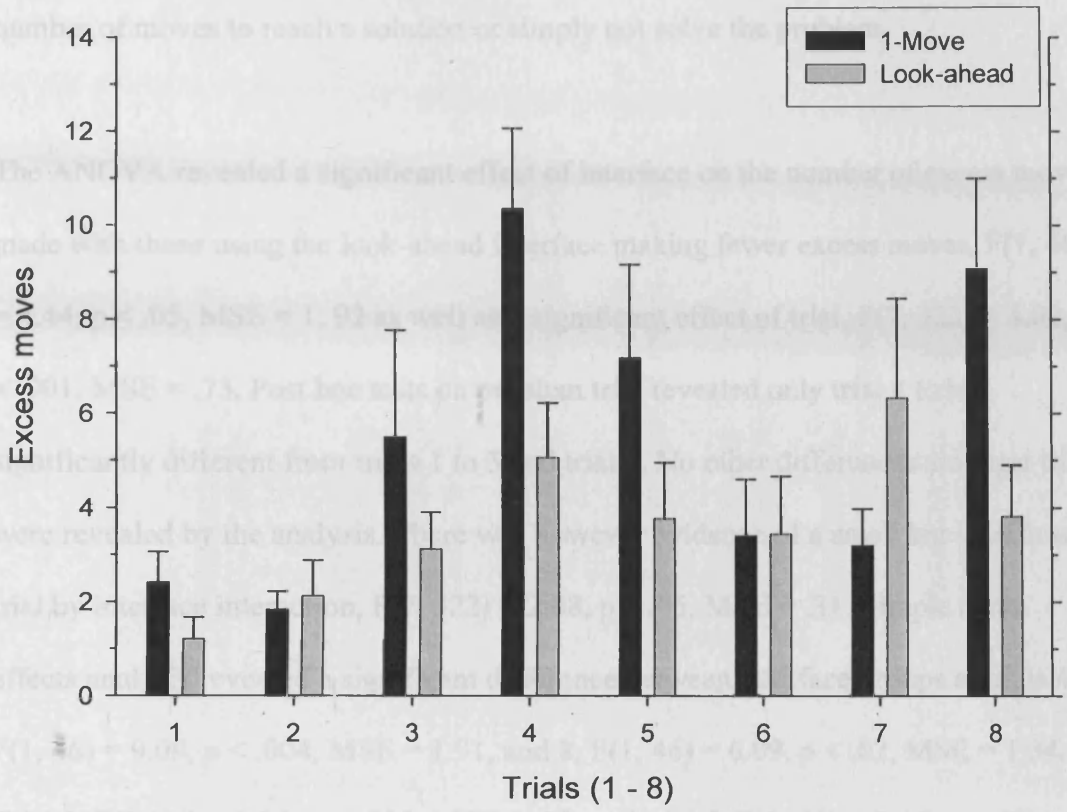


Figure 35. Number of Excess Moves made Over Trials by 1-Move and Look-ahead Interface Users

Error bars represent plus one standard error

From the data above it appears that trials 1 and 2, requiring only 3 moves to solution were solved with very little difficulty or excess numbers of moves. A further examination reveals that a ceiling effect on performance may have been reached by many participants on both these trials. Trial 6, requiring 5 moves to solution appears to have been solved with almost equal efficiency by both groups of participants. Trial 7 requiring 6 moves was solved using greater numbers of excess moves by look-ahead users on average. An examination of the proportion of participants solving trial 7 in the minimum number found this to be greatest with participants using the look-ahead interface. It therefore appears that those not solving problem 7 either with the optimal

path or a small number of moves soon after went on to either eventually take a large number of moves to reach a solution or simply not solve the problem.

The ANOVA revealed a significant effect of interface on the number of excess moves made with those using the look-ahead interface making fewer excess moves, $F(1, 46) = 4.44$, $p < .05$, MSE = 1. 92 as well as a significant effect of trial, $F(7, 322) = 4.02$, $p < .001$, MSE = .73. Post hoc tests on problem trial revealed only trial 4 to be significantly different from trials 1 to 3 and trial 6. No other differences amongst trials were revealed by the analysis. There was however evidence of a small but significant trial by interface interaction, $F(7, 322) = 2.08$, $p < .05$, MSE = .31. Simple main effects analysis revealed a significant difference between interface groups at trials 4, $F(1, 46) = 9.09$, $p < .004$, MSE = 1.91, and 8, $F(1, 46) = 6.09$, $p < .02$, MSE = 1.34. Trial 1, $F(1, 46) = 3.14$, $p = .083$, MSE = .39, and trial 5, $F(1, 46) = 2.46$, $p = .12$, MSE = .54, revealed slight trends but these were not statistically significant. All other trials revealed no evidence of any significant differences between groups (F's < 1).

Total Time

Total solution time for both interface groups can be seen below in Figure 36. Consistent with previous findings of look-ahead, no difference in total time to complete trials between interface groups was found, $F(1, 46) = 2.25$, $p > .1$, MSE = .45.

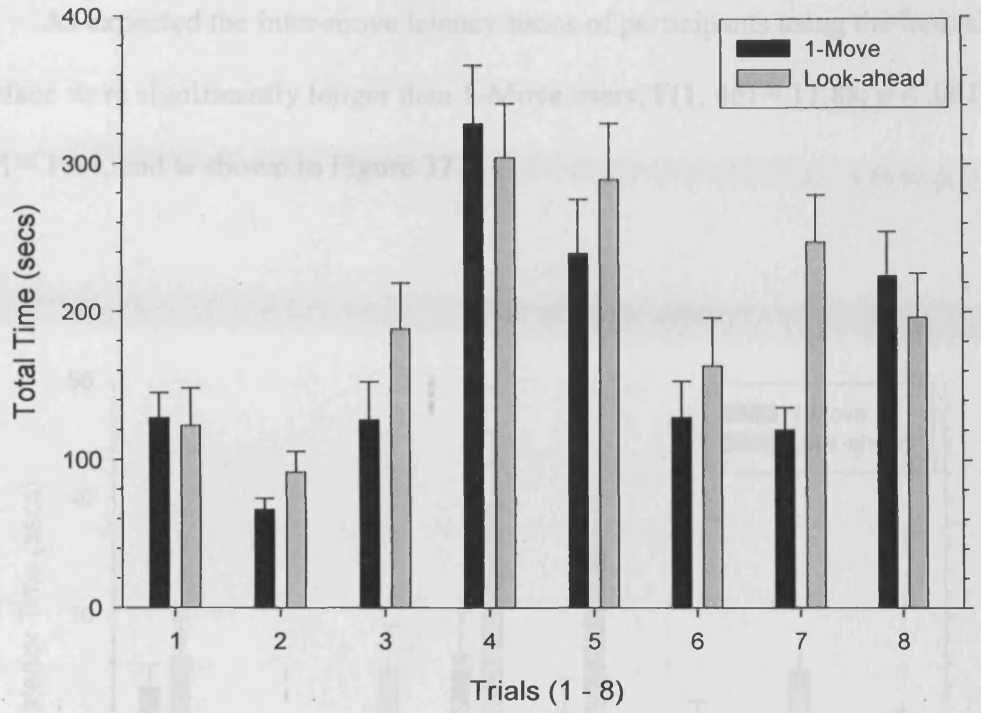


Figure 36. Total Time to Solution Over Trials for 1-Move and Look-ahead Interface Users

Error bars represent plus one standard error

The analysis revealed a significant effect of problem trial, $F(7, 322) = 15.39$, $p < .001$, MSE = 1.42, which was moderated by a small but significant trial by interface interaction, $F(7, 322) = 2.60$, $p < .02$, MSE = .24. Simple main effects revealed that trial 7 took significantly longer to solve when using the look-ahead interface, $F(1, 46) = 14.37$, $p < .001$, MSE = .98. Total time to solution between interface groups on any of the other problems did not reach significance (All p's > .1).

Inter-move Latency

As expected the inter-move latency times of participants using the look-ahead interface were significantly longer than 1-Move users, $F(1, 46) = 11.88$, $p < .001$, MSE = 1.64, and is shown in Figure 37.

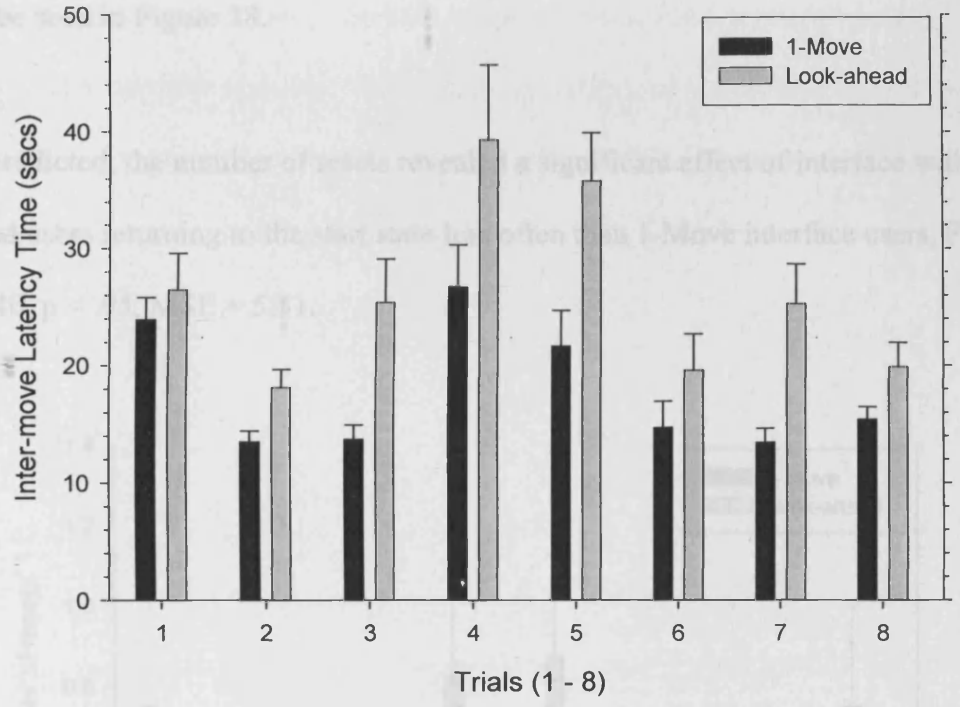


Figure 37. Inter-move Latency Times Over Trials for 1-Move and Look-ahead Interface Users

Error bars represent plus one standard error

The analysis revealed a significant effect of trial, $F(7, 322) = 10.95$, $p < .001$, MSE = .39. Post hoc tests by trial revealed trial 1 to be significantly different from trial 2, 3, 6 (p's < .01) and 8 ($p < .05$). Trial 2 was significantly different from trial's 4 and 5 ($p < .01$), while trial 3 was also different from trial 4 ($p < .001$). Trial 4 was significantly different from trial's 6, 7 and 8 (p's < .001) as was trial 5 (all p's < .01). No other

differences were significant. There was some evidence of a small trial by interface interaction, although it did not reach significance, $F(7, 322) = 1.86$, $p > .07$, MSE = .07.

Number of Resets

The effect of interface on the number of resets made to the original start sate can be seen in Figure 38.

As predicted, the number of resets revealed a significant effect of interface with look-ahead users returning to the start state less often than 1-Move interface users, $F(1, 46) = 4.40$, $p < .05$, MSE = 5.51.

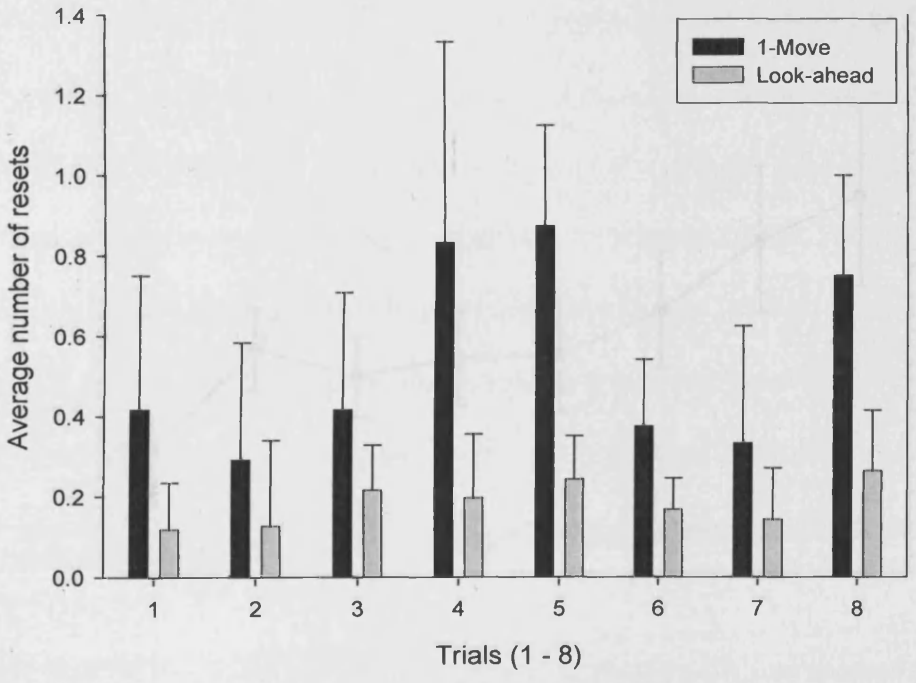


Figure 38. Number of Resets per Trial for 1-Move and Look-ahead Interface Users

Error bars represent plus one standard error

The analysis revealed no effect of trial, $F(7, 301) = 1.68$, $p > .1$, $MSE = .98$, and no evidence of an interface by trial interaction ($F < 1$).

Look-ahead Span

The numbers of moves specified per pour were examined and are presented in Figure 39 below. Interpretation of the data is made more complicated by the increasing differences in solution path length over trials and hence the range of look-ahead that is actually possible. The increasing number of moves may in some ways also act as a support by gradually requiring more look-ahead to be undertaken, so any conclusions must be drawn with caution.

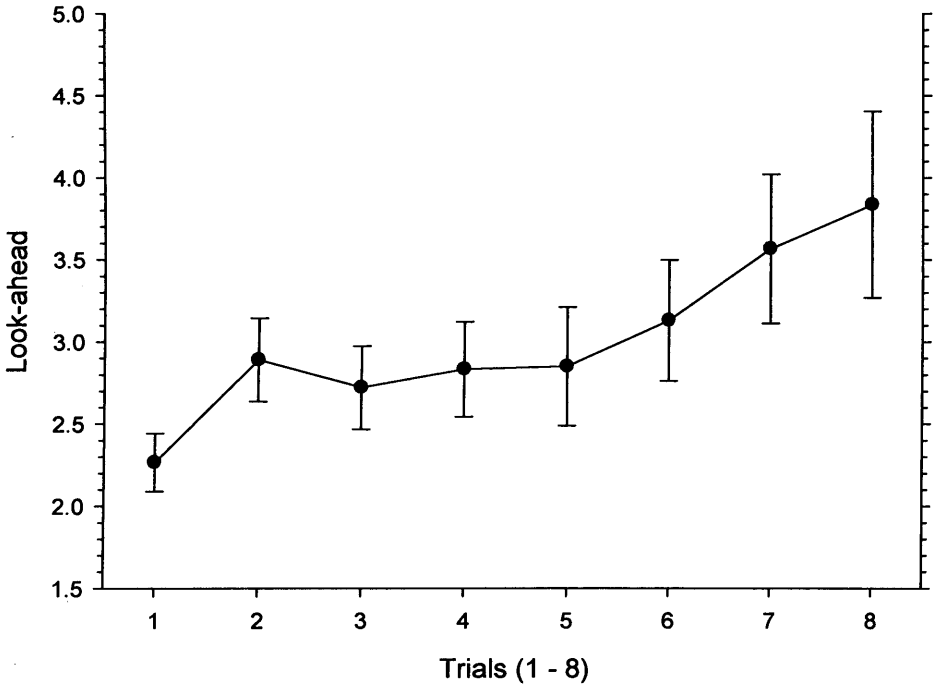


Figure 39. Mean Look-ahead Span Over Trials for Look-ahead Interface Users Error Bars Represent plus and minus one Standard Error

The results appear to replicate the previous 8-puzzle data regarding look-ahead span in terms of a linear increase in look-ahead over trials suggesting a greater proficiency with greater exposure to the task. A repeated measures ANOVA on look-ahead span revealed a significant effect of trial, $F(7, 161) = 3.41$, $p < .002$, $MSE = .08$. Trend analysis also revealed a highly significant linear component to the curve, $F(1, 23) = 8.25$, $p < .009$, $MSE = .414$). There was no evidence of any quadratic component to the curve ($F < 1$).

To further investigate the look-ahead process the number of pours (i.e. score per trial) of participants who successfully solved a trial was examined further. From the total number of solvers per trial, the proportions of participants who solved the puzzle by specifying 3, 4, 5, 6, or over 6 moves at once is given. For example, a high proportion of 1-Pour solvers on any trial would indicate that the solution path for that trial was discovered by a large number of participants and all moves required to solution could be specified in one attempt. The further the breakdown goes the clearer the observation is for how participants were actually solving problems. A large proportion of subjects solving problems in two pours for example may indicate a two-stage process in the solution. Table 4 below provides a breakdown over the 8 trials for participants who completed trials and provides a finer analysis of performance and given an indication of individual differences during the planning of solutions.

A consistent finding from Table 4 below is that approximately one third of all users in the look-ahead condition implemented the entire solution at once. The consistency of this finding, irrespective of solution length, may indicate that a proportion of participants were using a problem solving strategy that could guarantee strong

performance given any problem. Delaney et al. (2004) found evidence from verbal protocols that high planners were using an algebraic strategy to calculate solution paths. It may be possible that such a strategy was also being employed by a certain number of participants in the current experiment to achieve such high performance. Approximately one fifth of participants, excluding trial 7, used two pours to solve a number of the problems. This rate was higher for trials 3 and 4, both taking 4-moves to solution, perhaps indicating a two-stage process in the development and implementation of a plan. As has been previously argued (e.g. Greeno, 1974), a multi-step sequence of moves may be the conceptualisation of water jars performance. The data below seem to indicate that the use multi-step sequences may be appropriate, at least for a large proportion of participants.

Table 4.

Proportion of Solvers Reaching a Solution in Number of Pour Presses

	Problem Trial							
	1	2	3	4	5	6	7	8
1 Pour	.5	.583	.375	.368	.285	.428	.375	.375
2 Pours	.083	.125	.208	.210	.190	.142	.083	.187
3 Pours	.25	.041	0	0	.095	.047	.125	.125
4 Pours	.041	.125	.083	.157	.047	0	0	.0625
5 Pours	.041	0	.041	.052	.142	.095	0	.0625
6 Pours	.041	0	0	0	0	0	.041	.0625
> 6 Pours	.041	.125	.291	.210	.238	.285	.375	.125

While the look-ahead interface aimed to encourage the completion of a trial in as few pour presses as possible, there were instances of look-ahead participants recording a

score that was also consistent with the minimum number of moves for that particular trial. In other words, there was evidence that some trials were being completed by possibly entering only one move at a time, yet still completing the trials optimally. Examining the data revealed only 11 instances of trials having been solved optimally by entering in the required moves one at a time. From the 173 trials successfully solved is a proportion of only .063, indicating that it accounts for a very small amount of observed performance for look-ahead users and of optimal performance.

Discussion

In line with predictions, the look-ahead manipulation was successful in increasing performance along a number of key performance measures. A significantly greater number of trials were solved in the minimum number of moves. Similarly, Look-ahead participants required significantly fewer numbers of extra moves above the minimum to reach solution. This increase in observed performance however, did not result in fewer numbers of problems being solved, with both interface groups showing almost identical problem completion rates. Therefore, the probability of completing a trial by Look-ahead participant was not adversely affected by possible fixations upon achieving optimal or near optimal trial performance. This result would also indicate that increased planning and look-ahead may not necessarily increase the likelihood of problem completion but simply the chance that the problem will be solved with much greater efficiency.

Consistent with the results reported in the previous chapter, total solution time was generally not affected by the interface manipulation. It would appear that having to extract numerous pieces of information to decide on the next best move can be done by participants, even with ever increasing depth, yet leave total time unaffected.

If total time had been significantly longer for planning participants this result may have been due to the requirement to process multiple sources of information over a number of moves. The fact that no difference was generally found, except on one trial, would suggest that the multiple sources of information for move data are collapsed into a singular piece of information regarding the operator, represented in terms of the effect on the current states, so that depth of search can take place. Total solution time data therefore indicates that general problem solving efficiency can increase, but without increased costs in terms of time to solution. This is improvement is brought about by the extra time spent planning, indicated by significant differences in inter-move latencies and fewer excess moves. 1-Move performance can be better characterized as adopting less planning that leads to the implementation of superfluous or inappropriate moves or the adoption of a weak strategy to help generate the required solution path, as indicated by greater number of resets.

In addition to the above data, a number of other measures such as number of errors in proposed planned sequences of moves were at floor level. Even with increasing numbers of moves required to solution, plan accuracy appears to remain largely unaffected. Along with the observed high levels of accuracy, it suggests that formulated plans are precise and may support the argument about the natural progressive cost of ever deepening search. The cost of constructing an entire solution path in the water jars problem would be a process requiring large cognitive resources. Therefore, the decision to specify a move of poorer quality or that has not been thoroughly checked would be unlikely for look-ahead users given the difficulty in specifying a chain of meaningful legal moves. The fewer number of resets to the start state by look-ahead users also support the argument of less reliance on breadth of search for promising states being undertaken but rather depth of search occurring

from a possibly more advantageous first move. While 1-Move users appear to be adopting a more "try it and see" approach to solution path generation with increased numbers of resets as a result, Look-ahead participants appear not to rely on the resetting of states. This could also be taken as indirect evidence that they are performing and manipulating moves in the internal problem space that has been constructed of the problem rather than upon the external environment.

Look-ahead span measures of performance bear a striking resemblance to the results from Experiments 3 and 4, with both a linear increase in look-ahead occurring and depth of span being three to four moves by the end of trials. What is interesting is that given the greater effort required to gather information about the current states and the possible future states that may be entered into, look-ahead span remains at approximately 3 to 4 steps. This may indicate a general level of look-ahead span that remains relatively robust against task demands and characteristics, even when at more complex levels. The results may give a reasonable expectation of the span that novice participants may have when in a relatively novel domain. A further breakdown of look-ahead span also revealed that around one third of participants were solving the water jars problems through the specification of a complete plan, even when solutions required six moves in total. The remaining participants appear to be solving the puzzles using a variety of plan-lengths with approximately one fifth also solving puzzles using two pour presses. This may indicate a strategy for reducing working memory load by implementing the currently specified moves, so as to update the current contents of all three jars, before resuming another stage of planning. This two-stage step may of course be simply coincidence in terms of solution success, yet it is likely such a strategy served a particular purpose by leaving the solution what the

participant believed to be a promising state before using the second pour to specify any remaining move(s).

The current results are in line with those of Delaney et al. (2004) who also reported similar benefits of planning on water jars problems in terms of fewer excess moves to solution. The current findings also have implications for previous models of performance regarding the likelihood of participants considering any number of moves beyond a depth of one in the solution path (e.g. Atwood & Polson, 1976). A certain proportion of 1-Move users also performed in what appears to be a largely planful manner, as indicated by the total number of problems solved optimally. A number of factors may explain such a result. Firstly, the use of the Scoreboard manipulation may have restrained control performance to a certain degree. Given the small numbers of moves involved from start to end states, it could quite possibly have made participants more cautious in their move selection choice. The small increase in motivation to limit the number of moves made may have been enough to ensure that some control participants performed well beyond the norm. A second factor relates more to the current method of move selection during the problem solving process. The number of steps required to specify a move, compared with the ease of move selection in the 8-puzzle for example, may have also had an influence on the likelihood to act spontaneously. Previous cost based operator manipulations have shown that only a small associated cost is needed to make behaviour more planful (O'Hara & Payne, 1998). The slightly cumbersome specification process of choosing which contents to pour from jar to jar may also have lead to an adaptation by some control participants to restrain move selection. Look-ahead users may also of course

have been influenced in a similar manner as well as by the means by which to specify greater numbers of moves.

While the results provide support for the use of the current manipulation a number of issues remain unexplored. Firstly, there was evidence of ceiling effects on certain water jar problems in the current study. Therefore, a number of changes could be made to further clarify and extend the existing findings. The first is in relation to the number of participants being excluded through the screener process. Increasing the number of participants entering into the Water Jars phase may allow greater variability in performance with which performance differences can be more easily detected. The second change relates specifically to the use of the current Scoreboard intervention. While it has proven successful, there remain a number of other possible manipulations that could be of equal benefit and that are also more directly linked to the interaction between participant and interface used. These issues were more thoroughly examined in Experiment 6.

Experiment 6

Experiment 5 demonstrated that look-ahead can be encouraged, at the user's discretion, by the implementation of an external mechanism aimed at increasing performance to something more akin at times to expert performance. Whilst external or supplementary interface 'tools' may have benefits for performance and implications for areas like educational software development, it may not always be so appropriate in a more applied setting to provide such motivational aids during task performance. What may be more likely and indeed applicable is an interactive mechanism that is an integral part of the operational environment that users are immersed within.

Therefore, the current experiment had the express aim of confirming the positive effects of the previous successful manipulation to increase look-ahead and performance yet with an intervention that is intrinsically linked to the user interface, as opposed to external motivational aids such as a scoring system. A wide and varied number of approaches to interface manipulation have lead to the adoption of previously well established artefacts from human-machine interaction and used them successfully in recent problem solving manipulations. Prime among these approaches has been the manipulation of *system response time* (SRT) or *lockout time,* to examine the effects upon user evaluations, expectations, performance and behaviour with a particular interface. The two terms, although similar in nature, operate differently and a clearer description of each would be of benefit before proceeding. SRT would for example, be measured by the length of time that a system takes to respond to a user query for information. The typical example in today's world is someone's request for

information from the internet and the time taken from the user clicking a link on a webpage to the time it takes for the page to load. Lockout time differs primarily in the actual placement where the delay occurs. The input device will usually accept, process and display the information almost immediately. The lockout time will then be the interval between this point and the time that the next input can be entered or accepted by the system (Corley, 1976).

The issue of SRT's and lockout time are perhaps more relevant in our growing current technological environment than ever before and have become topical issues given the increasing amount and uptake of new internet technologies that require speed of response as a key requirement if they are to be successful. It is important to note that the *nature* of the task at hand will determine the applicability of such a proposed manipulation. For example, the introduction of a slow SRT on an internet website aimed at shoppers would most likely only have detrimental effects. Indeed recent research has shown that long SRT's on webpages lead to increased levels of stress (Trimmel, Meixner-Pendleton & Haring, 2003; Emurian, 1991, Guynes, 1988), negative ratings of content and experience (Ramsay, Barbesi & Preece, 1998), can lead to a loss of business or negative impressions of a company (Rose, Lees & Meuter, 2001) and a loss of income (Zona Research, 2001).

For over 30 years the impact of SRT on user performance has been acknowledged. It was the description, based largely on personal observation and experience that lead Miller (1968) to propose that system response times should be seen as analogous with typical human conversational processes. A 2-second delay time limit was the proposed maximum SRT that would still allow for users to feel in control of a system

without negative consequences. For optimum conversational flow and therefore a positive flow from humans to systems a time of around 0.5 seconds was argued to be the optimal SRT, a guideline that is still largely followed today (e.g. Nielsen, 1994). Lambert (1984) found that with increased responses, defined as sub-second response times, a programmer's productivity increased by an estimated by 62%. Other research (e.g. Butler, 1983) investigated the effect of system delays upon subjects' ability to complete typing and data entry tasks. No negative impact upon user performance was found on typical measures such as task accuracy or mean user typing time. Although variable system delays did show some evidence of affecting user response times the results were mixed and the effect was not strong. Other research (e.g. Martin & Corl, 1986), has found mainly negative effects of delays on problem solving performance but not on more real world tasks. The topic remains ambiguous and open to large differences in results when other factors are at play. For example, the use of percentage bars or other visual and audio aids have been shown to alleviate the negative impressions that system delays generally cause (e.g. Myers, 1985; Crease & Brewster, 1998) and a multitude of explanations can be put forward to explain the discrepancies that exist in the literature. Teal and Rudnicky (1992) provide an overview of some of the possible reasons ranging from differences in the choice of dependent variables, variable amounts of task delay settings and the nature of the task itself.

SRT's and lockout times, while typically viewed as unwanted, are not necessarily negative especially in an area where the *quality* of performance is the index of measurement rather than simply '*throughput*' or speed (Carroll & Rosson, 1987). When the yardstick of performance is learning for example, the potential benefits of

intentionally placed delays become more evident. Stokes, Halcomb & Slovacek (1988) assigned students sitting a questionnaire test via a computer to either a delay group or non-delay group. Students in the delay group were presented with the questions, upon which a lockout time was triggered and were prevented from immediately answering the questions until after a period of time. This simple measure was enough to significantly increase performance from controls and suggest possible implications for the implementation of online learning and testing materials.

In a more typical problem solving task Grossberg, Wiesen & Yntema (1976) found that the use of a delay changed the approach adopted by participants with evidence of more careful and calculated choices being made which subsequently led to fewer errors and more efficient solutions. A difference in the choice of problem solving strategy may be one particular outcome of a delay mechanism. Support for such an argument comes from Child (1999) who examined the impact that a 5-second delay had upon the information location strategies of users with a specialized hypermedia system when compared to a standard non-delay group. Using the hypertext environment participants were required to find information on three different topics. The effect of the 5-second delay produced significant differences between the groups. The non-delay group switched pages more frequently, took much longer to complete tasks, accessed greater numbers of total pages and had to backtrack more often than the delay group. Child (1999) attributed these differences to the effect that the delay had upon the subjects' choice of problem solving strategy. Analysis of the behaviour revealed evidence of a "locate-in-breath" strategy for those completing the task without a delay while a "locate-in-depth" strategy was evoked in users affected by the delay (Child, 1999). These users tended to examine greater numbers of pages located

in their sequential ordering within the same section. Interestingly the delay affected users irrespective of amount of previous experience or skill with the hypertext environment, suggesting that such a manipulation can be applied to all users while still allowing flexibility.

More recently O' Hara & Payne (1998) specifically manipulated lockout time and its subsequent effect on problem solving performance using the 8-puzzle. Control subjects, with a 3.3 -second lockout time, took larger numbers of moves to solve trials than those with a 7 second lockout time. The problem start state however remained constant throughout all trials which may have contributed somewhat to the effect. It is unclear if such a mechanism would be as successful with constantly changing start and goal states. Similar to other studies of planning reported here total time to solution was unaffected, with the completion of trials with greater efficiency compensating for the increased delays caused by the lockout time.

Any lockout time would most likely only prove effective in the current experiment if set to an unusually large period of time. SRT was chosen due to the long inter-move latency times already evident in water jars problems, indicating that users would be somewhat immune to any lockout time manipulation as they would simply use the time planning, bypassing the intended inconvenience of a lockout-time. The planning groups in the O'Hara & Payne study may have been become somewhat immune to any effects of lockout time, as it allowed the continuation of planning rather than obstruct initial performance at least, in any negative way. With the current manipulation, performance is somewhat more disrupted as the current contents take a period of time to be confirmed and may prohibit any extensive future planning if

planning is only undertaken in very small amounts. In the same way users may have adapted to the lockout times in the O'Hara & Payne (1998) study, there of course may be a similar adaptation by participants in the current experiment. Participants may become largely unaffected by the SRT if they have completely pre-planned entire solution paths. However, from the previous experimental findings this may only apply to a certain proportion of subjects while still affecting a significant proportion of the remaining population.

A SRT time manipulation may have more of an impact, as users will be forced to wait for visual confirmation of the effect their proposed move(s) had upon the current contents of the Water Jars. This should significantly affect their ability to begin planning any next move(s), ensuring the SRT manipulation is having an effect. While the longer the SRT, the greater the likelihood of an increase in planning, it would detract somewhat from the applicability of the current manipulation to any other field of study. Therefore a delay of only 4 seconds was selected as the SRT, with the motivation to look-ahead left entirely to the interaction caused by the delay between the problem and participant.

Predictions of performance are somewhat difficult given a number of new factors. The first of these involves simplifying the interface by making it easier to specify moves. Previous manipulations on operator cost (e.g. O'Hara & Payne, 1998; Golightly & Gilmore, 1997; Golightly, 1996) have shown that when the cost of implementing change is high the planfulness of behaviour also increases. The previous three step system for implementing moves may have been somewhat burdensome and as an unintentional by-product increased the amount of planning partaken by participants in both interface groups. Reducing the number of steps

required to specify a move may change the approach adopted by participants and mirror the ease of interaction that most interfaces aim to achieve in current environments.

The second adjustment is in regards to the difficulty of the screener that participants complete prior to their acceptance into the water jars phase of the experiment. A large proportion of participants (.46), were screened out in Experiment 5 after failing to answer the necessary number of questions correctly. The difficulty of the screener may have prohibited a more representative sample of the population taking part and as a result only permitted more capable participants through to the water jars phase of the experiment. Reducing the difficulty of the screener should ensure a more representative sample take part in the experiment.

The removal of the scoreboard system in the current experiment removes a purported motivation for users to perform more efficiently than they would do without such feedback. The performance observed from control subjects from the previous experiment may support such a conclusion. Given the effect that an external motivator may play in enhancing performance, the removal of such visual feedback may lead to more variable performance, as seen in Experiment 1, as participants feel less inhibited in move choice. However, Look-ahead interface users should still perform significantly better than controls, due to the lack of any inherent interface mechanisms to plan or external motivations to enhance performance for control subjects. There should therefore still be a significant number of trials solved in the minimum number of moves as well as fewer excess moves made overall.

The effects of increased performance should mirror previous findings from Experiments 5 and 6, in that no significant difference in the total time should be evident due to the increased performance invoked by the planning manipulation. It

should also be expected that those experiencing a SRT delay will make fewer resets than controls in accordance with the previous findings from Experiment 5.

Specific predictions for look-ahead users remain difficult to predict. Given the small SRT delay of only 4-seconds incurred through pressing the 'Pour' button, it may be that look-ahead span may change in the current experiment. If the SRT mechanism works as intended then it should ensure greater look-ahead as the natural costs inherent in the problem remain unchanged by the manipulation. If this is the case then similar levels of look-ahead span, as in Experiment 5, should still be observed with a gradual increase over trials. Alternatively, if the Scoreboard manipulation was effective in helping to eliminate wasteful moves, then its absence may have the effect of increasing the numbers of moves participants are willing to make as no visible penalty in terms of feedback of performance is available. If participants judge the current manipulation to be of less weighting, given the small time cost incurred, then look-ahead span may change as a direct consequence. With only a small time delay to restrain and influence problem solving, an adaptation in terms of behaviour may be likely. It would be both reasonable and predicted that a greater time delay of 10 seconds for example would raise the performance accordingly. Yet from a practical point of view this would appear too severe and reduce the applicability of the current manipulation for possible future work and in other related areas. If a small time delay can increase performance while still remaining applicable in a general sense then the smaller delay time would be more informative for future manipulations.

Analysis of excess moves in Experiment 5 indicated ceiling effects on certain problems and hence caution was warranted when making firm conclusions. The 3-

move problems in particular were prone to high performance by all participants. It may prove that in tandem with a simple scoreboard system and given the small numbers of moves to solution that the problems require, 1-Move performance can actually be increased sufficiently by such a simple intervention and that planning adds little extra. Such a conclusion however may be premature and a number of small changes to the design and materials may be enough to increase the variability in performance and reduce the likelihood of ceiling effects. It would therefore be predicted that larger difference in excess moves between interface groups would be observed in the current experiment.

Method

Participants

A total of 52 Cardiff University students took part for a payment of £6 or course credit. None of the participants had taken part in the previous experiments or reported having any experience with water jars problems. All participants were given the screener from experiment 2 of which 12 failed to answer two or more questions correctly and were excluded from the experiment. The remaining 40 participants (Mean age = 21.95 years, S.D. = 2.65) were randomly assigned to one of two interfaces with 20 participants in each condition.

Design

The between subject manipulation of interface had two levels with a 1-Move control group or a SRT Look-ahead group. The within subject variable of trial had six levels, details of which can be found in Appendix F. As in the previous experiments trial completion rates, proportions of trials solved optimally, number of excess moves

per trial, total solution time and number of resets were recorded for all participants. Inter-move latency times were not calculated due to the possible confound of SRT. For those in the look-ahead condition look-ahead span, number of errors and proportions of solutions solved in number of pour presses were also recorded.

Materials

The presentation of water jar contents and method of move validation were identical to those used by both interface groups in Experiment 5.

Due to the indication of ceiling effects in the previous experiment a number of changes were made in Experiment 6. The screener's difficulty was reduced slightly to allow for a wider selection of participants to take part in the Water Jars phase. A copy of the questions used can be found in Appendix H. The second change simplified the move specification process, so that only one button press was required in order to specify a move in its entirety. An example of the interface that Look-ahead users were given is shown in Figure 40 below.

The process of move implementation operated in the same manner as previously, with any moves specified having to be validated in their entirety before any change to the scoreboard or current contents could take place. Participants were notified of any errors in moves specified by a message box, previous moves were removed from the list and they began again.

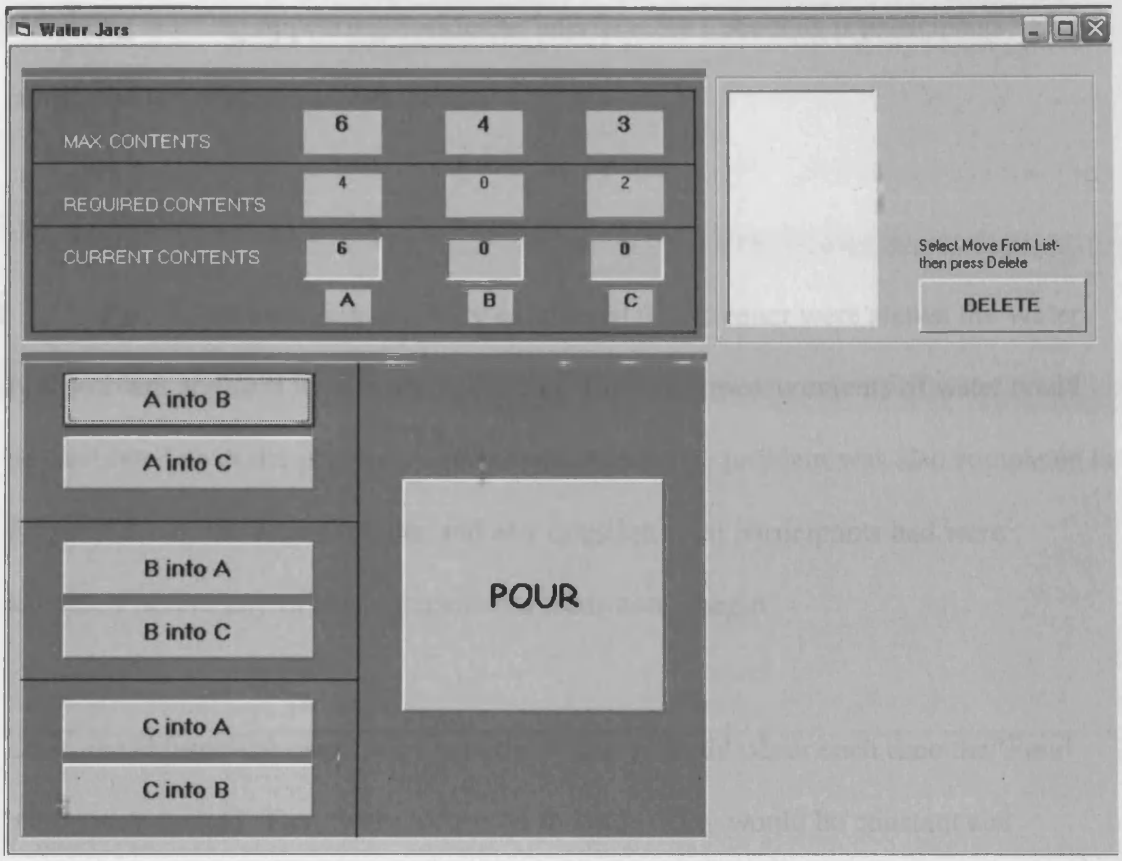


Figure 40. Look-ahead Interface Used in Experiment 6

Participants were simply instructed in the current experiment to solve each problem in as few pour presses as possible. For look-ahead interface users a 4-second SRT time was introduced whereby the current contents would not update or the interface allow further user input until the delay had elapsed. The SRT was triggered by the pressing of the 'Pour' button upon which the system would remain unchanged and would then implement the move(s) that participants had specified. Participants would also only be made aware of any errors in their moves once the delay had elapsed.

The water jars problems consisted of 6 trials with one practice trial and five experimental trials that varied in the number of moves required to solution from 4 moves to 6. A 10 minute time limit was imposed for each trial, with a 3 minute time

remaining warning appearing beside the interface for 6 seconds if participants had not completed the trial within 7 minutes.

Procedure

Participants who successfully completed the screener were shown the Water Jars problem and told its aim and the means by which measurements of water could be calculated as in the previous experiment. A practice problem was also completed in the presence of the experimenter and any questions that participants had were answered before any of the experimental trials could begin.

Look-ahead interface users were told that a delay would occur each time the 'Pour' button was clicked. They were informed that this delay would be constant and throughout all trials. It however was triggered by the pressing of the pour button irrespective of the number of moves they had entered and that they could decide upon the number of moves to specify if they felt it would be of benefit to their progress. The instructions from the previous experiment, in regards to the programs response to errors in proposed plans, were explained to participants. They were also informed that the interface and current contents would remain unchanged until a move(s) had been checked for legality.

Results

General Performance Data

Examining overall completion rates revealed that 84.5% of problems attempted were solved successfully. Examining number of successful completions by interface group revealed that 88% of trials were completed by 1-Move users while

82% of trials were completed by look-ahead interface users. A Chi-square test on the percentages of successful completions versus non-completed trials by interface group revealed this difference not to be significant ($\chi^2 = 1.41$, n.s., df = 1). A breakdown of trials successfully solved by interface group is shown in Table 5 below.

Table 5

Percentage of Trials Successfully Completed by Interface Group

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
1-Move	85%	90%	80%	95%	90%
Look-ahead	70%	85%	85%	75%	90%
Total	77.5%	87.5%	82.5%	85%	90%

The measure of number of errors during plan specification for Look-ahead users was once again at floor level, with the modal number of errors by participant and by trial remaining at zero.

Optimal Performance Data

From the 100 trials attempted by participants in each interface group there were a total of 32 trials solved in the minimum number of moves by look-ahead users. 1-Move interface users solved only 14 such trials in the minimum number of moves. A Chi square test on the total number of trials solved optimally versus those not solved optimally revealed a highly significant effect of interface ($\chi^2 = 9.15$, $p < .01$, df

= 1). These results are again consistent with the results regarding optimal performance from Experiment 5 for Look-ahead users.

Number of Excess Moves

Number of excess moves made, total time to solution and move latencies were analysed using a two-way mixed ANOVA with interface as the between subject factor and trial as the within subject factor. The number of excess moves, m, was $\log_{10}(m + 1)$ transformed for the purposes of analysis. Missing values were replaced with the grand trial mean of finished participants' data (cf. Delaney et al., 2004).

As predicted, there was a significant effect of interface on the number of excess moves made, $F(1, 38) = 13.81$, $p < .001$, $MSE = 3.97$. Figure 41 below, shows the number of excess moves made over trials.

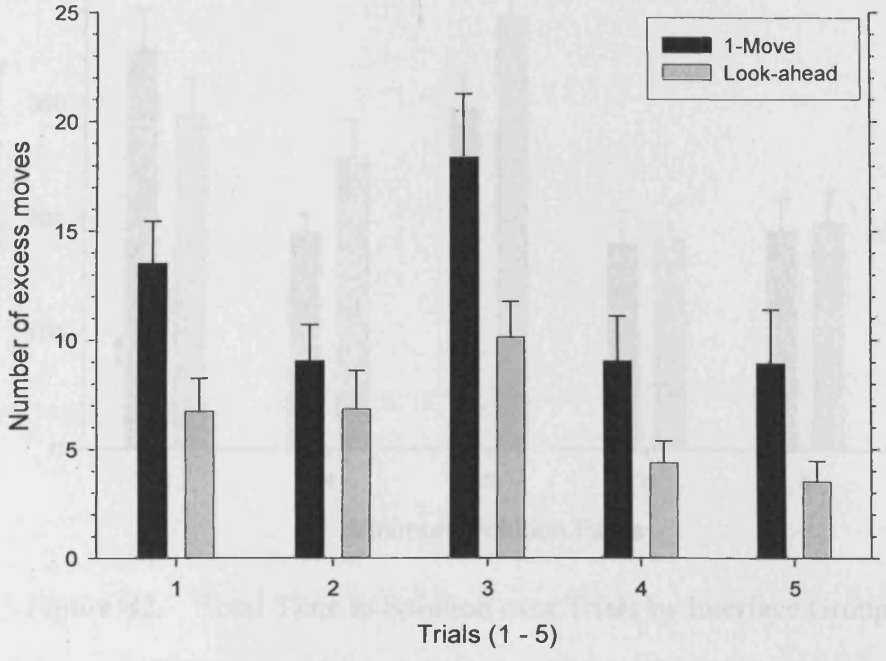


Figure 41. Number of Excess Moves Over Trials by Interface Group

Error Bars Represent Plus One Standard Error

The analysis also revealed a significant effect of trial, $F(4, 152) = 4.12$, $p < .003$, MSE = .79. Post hoc tests revealed a significant difference on the number of excess moves on trial 3 compared with those made in trials 2, 4 and 5. There was no evidence of any trial by interface interactions ($F < 1$). No other differences were found to be significant.

Total Time

In line with previous total time data, the ANOVA revealed a no effect of Interface on time taken to complete problems, $F(1, 38) = 1.38$, $p > .2$, MSE = .12. The analysis also revealed no significant trial by interface interaction, $F(4, 152) = 1.08$, $p > .3$, MSE = .07.

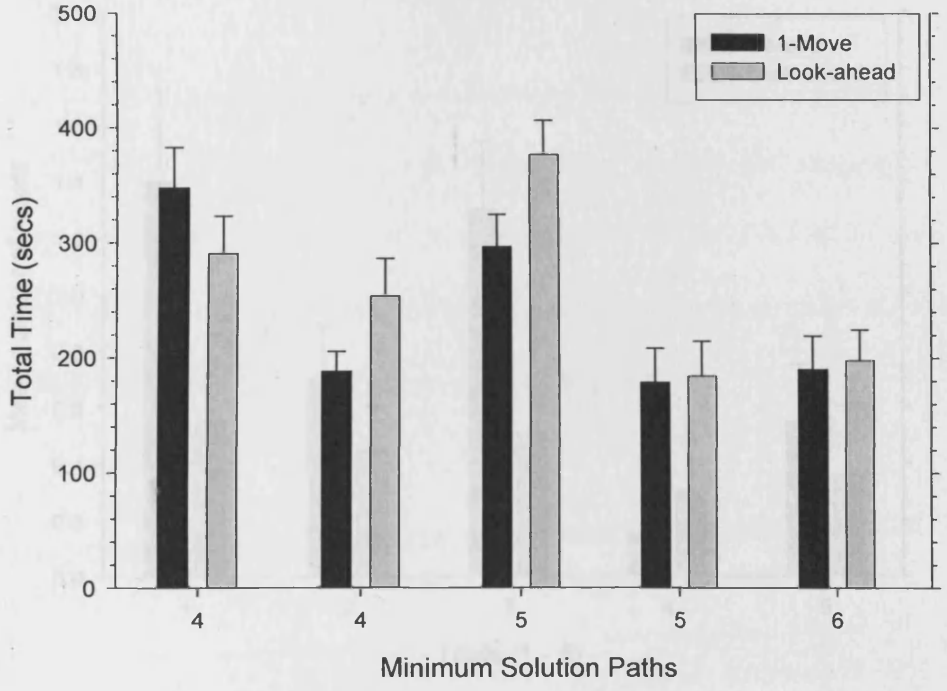


Figure 42. Total Time to Solution over Trials by Interface Group

Error Bars Represent Plus One Standard Error

The analysis did reveal a significant effect of trial, $F(4, 152) = 8.96$, $p < .001$, MSE = .61. Bonferroni corrected post hoc comparisons revealed that trial 1 took longer to solve than trials 2, 4 and 5. Trial 2 was significantly longer than trial 3 only. Trial 3 took longer than trial 4 and 5. No other differences were significant.

Number of Resets

As expected, there was a significant effect of interface on the number of resets to the start sate, with fewer resets being made by look-ahead users, $F(1, 38) = 4.84$, $p < .05$, MSE = 7.61. The result (see Figure 43 below), was modified by a small but significant interaction between trial and interface, $F(4, 152) = 2.49$, $p < .05$, MSE = 1.81.

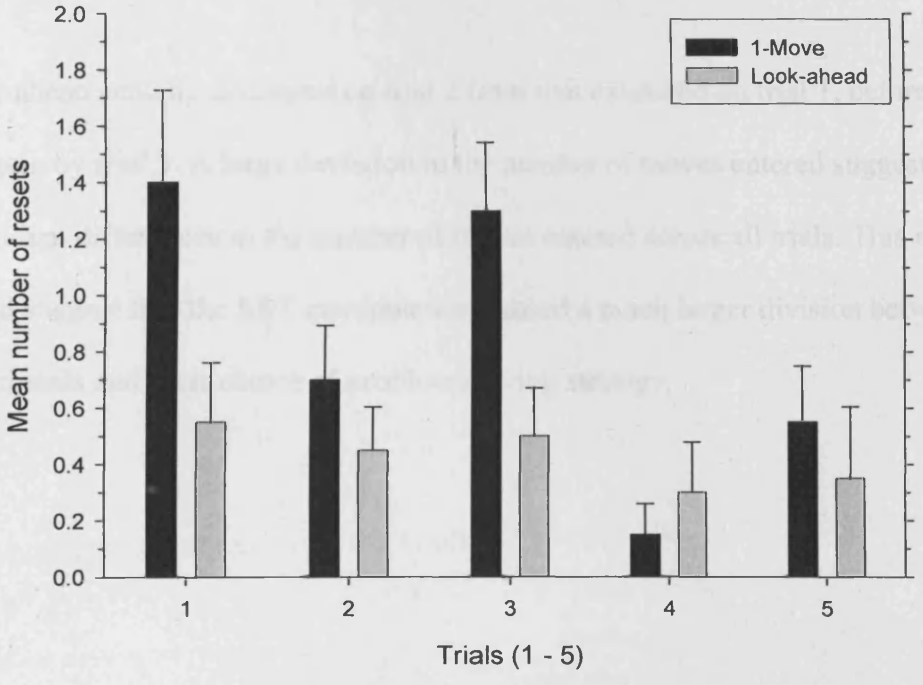


Figure 43. Number of Resets Made by Interface Users Over Trials

Error Bars Represent Plus One Standard Error

Simple main effects analysis revealed a significant difference at trial 1, $F(1, 38) = 4.75$, $p < .05$, MSE = 7.22, and at trial 3, $F(1, 38) = 7.32$, $p < .01$, MSE = 6.40. No other differences were significant (p's > .1). The analysis also revealed a significant effect of problem trial on the number of resets ($F(4, 152) = 5.37$, $p < .001$, MSE = 3.912). Post hoc comparisons on the number of resets made by trial revealed that fewer resets were made on trial 4 compared with trial 1 ($p < .02$) and trial 3 ($p < .002$). No other differences were significant.

Look-ahead Span

The results, shown in Figure 44, are in contrast to previous findings in that they do not show the clear linear increase in adding more moves to be specified at once.

Look-ahead actually decreases on trial 2 from that exhibited on trial 1, before picking up again by trial 5. A large deviation in the number of moves entered suggests large individual differences in the number of moves entered across all trials. This result would suggest that the SRT manipulation caused a much larger division between participants and their choice of problem solving strategy.

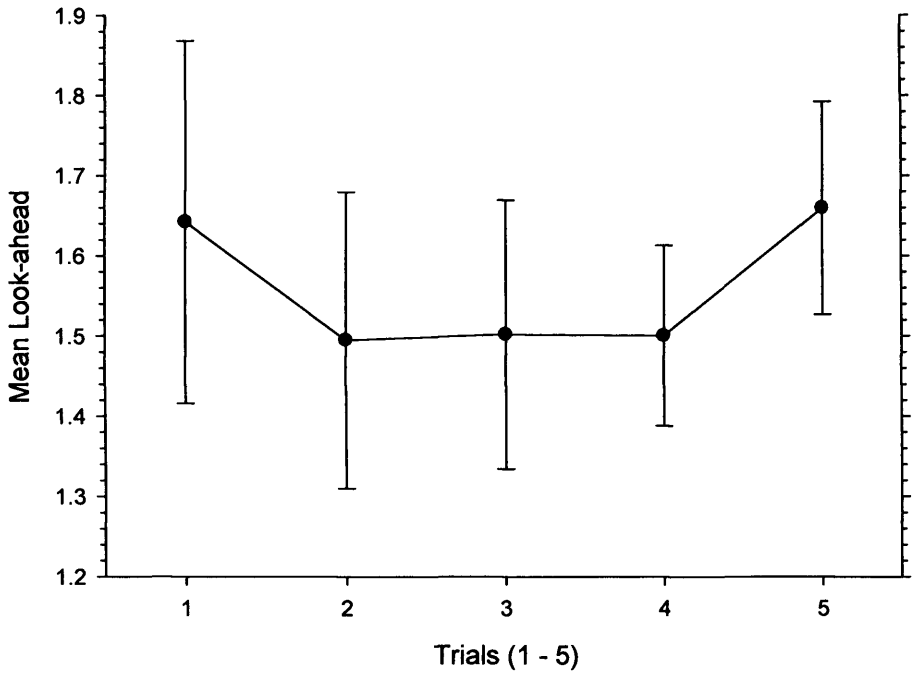


Figure 44. Average Look-ahead Span Over Trials by Look-ahead Interface Users Error Bars Represent Plus One Standard Error

The previous experiments have clearly indicated that given the correct motivation to plan participants can increase the amount of look-ahead undertaken. The lack of a pattern as found in previous work would suggest the utility of planning in the current circumstances was of less necessity.

Once again, proportions of pours used to generate solutions, were calculated for look-ahead users and are shown below in table 6. A larger proportion of trials were solved within 4 pour presses and apart from trial 1 and 2 performance the number of trials solved in one pour was actually zero. The results are very different to the data from experiment 5 with no participants having solved trials 3 to 5 by specifying the entire solution in one attempt.

Table 6

Proportion of trials solved in number of pours for Look-ahead users

	Trial				
	1	2	3	4	5
1 Pour	.142	.058	0	0	0
2 Pour	.071	.117	0	.133	.055
3 Pour	0	.176	.117	.133	0
4 Pour	.214	.176	0	.2	.388
5 Pour	.142	0	.176	.066	.166
6 Pour	.142	.117	.058	.066	.111
> 6 Pour	.285	.352	.647	.4	.277

This result in conjunction with the look-ahead span data further supports the assumption that the motivation to increase the number of moves entered in the current experiment was less powerful than that of the previous manipulation.

Discussion

The adaptive nature of human behaviour appears to have been exemplified by the current results when compared to those from Experiment 5. The very small delay in terms of SRT was predicted to have an effect on both the pattern of participant behaviour and performance while still increasing planfulness and efficient problem solving. In relation to the number of optimal solutions, Look-ahead users still completed one third of all trials in the minimum number of moves, well above the number completed optimally by 1-Move users.

For both interface groups, greater numbers of excess moves over trials were made when compared with the same performance measure as reported in Experiment 5. However, look-ahead users' performance in the current experiment still remained significantly better than control 1-Move performance. This reduction in performance was somewhat predicted, if the weighting given to the current manipulation was judged to be of lesser importance than the previous Scoreboard manipulation by participants. Therefore, performance would be a reflection of both the change in interface and participants' perceptions of the importance of the current manipulation to excel on trials if possible. Ericsson & Lehmann (1996) argue that subjects often demonstrate "maximal adaptation" in expert performance, determined largely by the interplay between continual training and natural ability. In a similar vein, current performance may have been maximally adaptive in terms of performance, predicted by natural differences in the ability to plan and the perceived impact of the current problem solving environment on participants' willingness to plan.

One potential consequence of this artefact would be a possible future examination of how behaviour adapts with increased SRT. It would be likely and predicted that planning and number of moves specified at once would increase as a reaction to longer SRT. This would also affect performance in terms of fewer excess moves and look-ahead span. There would undoubtedly be a time when a long or variable SRT will become counter productive (Long, 1976), yet there appears enough scope to allow for increasing the time used in the current experiment while still remaining a plausible mechanism in other interactive environments.

Total time to solution, one of the most robust findings throughout all the experiments reported, again showed no negative effects due to the planning manipulation. The

time taken to solve a problem appears to be dictated solely by the problem being solved rather than the interface being used in the current water jars experiments. It supports previous arguments for the applicability of planning interfaces given the appropriate context. They do not necessarily, if implemented correctly, lead to interfaces that detrimentally affect time taken to complete a task and allow greater application to related interface based domains where interaction is of vital importance.

Look-ahead participants did not demonstrate the same span of search as they had done previously, which is most likely again due to the shift in terms of the perceived benefit of doing so. If no clear benefit of specifying greater numbers of moves was perceivable to Look-ahead users then perhaps a more cautious approach was the main mechanism that led to increased performance rather than look-ahead per se. Knowles & Delaney (2005) demonstrated that increased caution due to increased cost in move choice was responsible for a reduction in the number of illegal moves made during problem solving. A similar effect in terms of increasing the time simply spent checking an intended move may be responsible for the current findings. Given the proportion of participants solving problems in 3 pours or fewer would suggest that a number of participants did increase look-ahead beyond the minimum. Increased planning was in operation but simply not being demonstrated through the actual specification of the pre-planned moves. Given the large number of optimal solutions and the current context it would be difficult to argue against at least a significant use of look-ahead in the current experiment.

The performance data exhibited by 1-Move users also appear to be of interest in the current experiment. It appears that in comparison with their performance from the previous experiment they were affected more detrimentally by the current changes

in design. The lack of feedback for performance appears to have resulted in a greater willingness to make greater number of moves when the restraints of a scoring system have been removed. This argument is pertinent to previous observations regarding the likely approach taken from control subjects in the Delaney et al. study. Without any restrictive practices in operation the chances of control participants exhibiting any constrained problem solving behaviour would be unlikely. Given the planful approach insisted upon by Delaney et al (2004), their plan group may be also more likely to adopt planful behaviour subsequent to any attempt to implement a plan that had perhaps gone wrong. The large effect in the current experiment is in some respects also due to the decrease in performance of the control group.

The final experiment of the thesis, again using the Water Jars problems, aimed to directly test for differences in performance between a total-order manipulation as used by Delaney et al. (2004), a Look-ahead or partial-plan group and a control group. The experiment once again adopted the 'Scoreboard' manipulation as implemented in Experiment 5.

Experiment 7

The results from Experiment 6 provide additional support for the current argument that increased levels of planning and look-ahead lead to improved problem solving performance, although the results were less consistent in terms of look-ahead span. However, pure measures of performance in terms of the number of trials solved in the minimum number of moves remained consistent, showing that look-ahead still lead to a greater number of optimal solutions. Once again total time was unaffected by the increased time spent planning.

There remains one particular question that remains largely unanswered and in many respects is one particular motivation for the current work. To what extent can look-ahead and planning be extended to and when, if ever, does this manipulation begin to become counter productive. Do the benefits of interface manipulations that simply encourage the participant to partake in more planning when possible out-perform a manipulation that demands either larger amounts of planning before action or complete plans to be specified? While existing evidence (e.g. Ratterman et al., 2001) indicates that total-order planning is often inadvisable if a complex subgoal structure is involved, evidence from recent Water Jars experiments have suggested that total-order planning is a viable means by which to solve a given problem when the task lacks such a structure. The enforcement of look-ahead in Experiment 1 of the current work also found enforced planning to be inadvisable. The problem with the enforcement seemed not to lie in the inability to plan 3-steps along a solution path but that such a strict span enforcement could not lead to the generation of a meaningful or efficient solution path, even after multiple attempts.

Although the benefits of planning, in terms of fewer excess moves, from the Delaney et al. (2004) study were informative, a number of key problem solving measures of performance were not reported by those authors that would for comparison purposes prove useful given the current findings. Firstly, performance data such as the number of trials solved in the minimum number of moves were not disclosed. There may have been a large discrepancy between problem solving accuracy and the participants' subjective report of being able to solve the problem. A particular point of interest is the relationship between the verbalisation that a plan has been formulated and its subsequent relationship to actual problem solving performance.

Secondly, no time data were reported in terms of either the time spent planning a solution (i.e. time taken to generate the entire solution), 'play time' (i.e. time taken to implement the planned moves; cf. Ward & Allport, 1997) or total time (planning time plus 'play time') to solution. A possible methodological concern is that participants were told to plan an entire solution without considering the very real possibility that a number of participants may not actually have been able to generate a solution. Subjects simply gave a verbal response to the experimenters that they were ready to begin entering their solution, without any attempt to check for participants' certainty they had in fact been able to generate the solution. No reports were given of participants being unable to plan a solution, yet it seems highly likely from both evidence in the literature and current findings that a certain number of subjects would have been unable to comply with the task requirements. The likelihood of this inability to fully calculate all the moves required may also be higher with problems that require longer solution paths although this, as discussed previously, is not certain.

The current experiment therefore aimed to compare control performance with two different planning groups. The Look-ahead manipulation as used in Experiment 5 and a slightly adapted version (see materials section) of the manipulation implemented by Delaney et al. (2004), known from here on as 'total-plan' users.

While enhanced performance is predicted for participants in both planning groups (look-ahead versus total-plan) when compared to 1-Move controls, differences that may arise between the two planning groups are much more difficult to predict. The main purpose of the current experiment is to test the effectiveness of both plan based manipulations. Being required to have the complete plan before beginning, may mean that total-plan participants solve a greater number of problems in the minimum number of moves compared to look-ahead users who are simply encouraged to solve puzzles in as few "moves" as possible, indicated by the Scoreboard trial score. Alternatively, if total-plan users' plans are inaccurate or a mistake is made during the implementation phase, total-plan users may either solve fewer numbers of problems in the minimum number of moves or be no different from Look-ahead participants. A second motivation for the current experiment was not only to test performance differences between plan manipulations, but also to test for differences in the ability to deal with solution paths of greater lengths in particular. The experiment tested for the ability of plan interface users to complete short solution path problems, both requiring 4 moves, versus longer solution path problems requiring 6 and 7 moves respectively.

While no difference may be predicted in performance on problems that take only a limited number of moves to solution, a difference between planning groups may be predicted on later problems that require increased numbers of moves to solution. The numbers of total-plan users indicating difficulty in forming an entire

pre-planned solution for problems requiring a large number of moves would also be predicted (see materials section for details).

Predictions for total time to solution data would most likely be that no difference between interface 1-Move and Look-ahead groups will be found, in accordance with previous findings. For Total-Plan users the predictions are more difficult to make. If users are able to formulate full-length solutions and not take an exponential time to do so with increased solution length, then no difference in total time would also be predicted. Look-ahead and total-plan users will once again use the extra time not spent acting superfluously upon the interface, planning more efficient solutions to compensate for any otherwise wasted time and therefore complete the problems in the same time as 1-Move users. If however, total-plan users find it difficult to generate an entire solution path for any of the problems then overall time to complete problems may be longer than controls and look-ahead users.

Method

Participants

A total of 70 Cardiff University students were recruited and given either course credit or a payment of £6. All participants were firstly screened with the four item mental arithmetic test as used in experiment 6 (see Appendix H). 19 Participants failed to answer 2 or more question correctly after two attempts per question and were removed from the experiment. The remaining 51 participants (Mean Age = 22.60 years, S.D. = 2.37) were randomly assigned to one of the three interface conditions with 17 participants in each experimental condition.

Design

The between subject factor of interface formed the basis of the experiment, with three levels: A control 1-Move condition, a Look-ahead condition and a Total-Plan based manipulation, which was a close replication of the planning manipulation implemented by Delaney et al. (2004). All three groups also had the Scoreboard present while solving problems. The within subject factor of trial had 4 levels with all water jars puzzles being presented in a fixed order (see Appendix G for details). Dependent measures of total time to solution, number of excess moves, number of successful completions of trials, number of trials solved in the minimum number of moves and number of resets during solutions were recorded for all participants. A look-ahead measurement was calculated for the look-ahead interface group as well as any errors in plans that could not be implemented. For the Total-Plan group additional measures of 'Plan' versus 'Cannot Plan' decisions made prior to move implementation was also recorded. Additionally, total time spent pre-planning and time spent implementing a solution (play time), were also recorded by the program.

Materials

There were 6 water jars problems in total, with 2 practice problems requiring only 3 moves to solution and 4 experimental trials. The 4 problem levels were chosen particularly with their solution path lengths in mind. The first two problems required only 4 moves to solution while problems 3 and 4 required 6 and 7 moves respectively. These were chosen to examine the applicability of total-plan manipulations with problems requiring increasing numbers of moves to solution compared to problems not requiring such extensive planning.

The Scoreboard system used in Experiment 5 was once again implemented in the current experiment for all users in conjunction with the interface used in Experiment 6. The total-plan group, similar to the condition used by Delaney et al. (2004), used the 1-Move interface to specify and implement their moves. Figure 45 below shows the screen that total-plan users encountered when a trial began. Two buttons were presented below the Scoreboard, one of which had to be pressed before move implementation could begin.

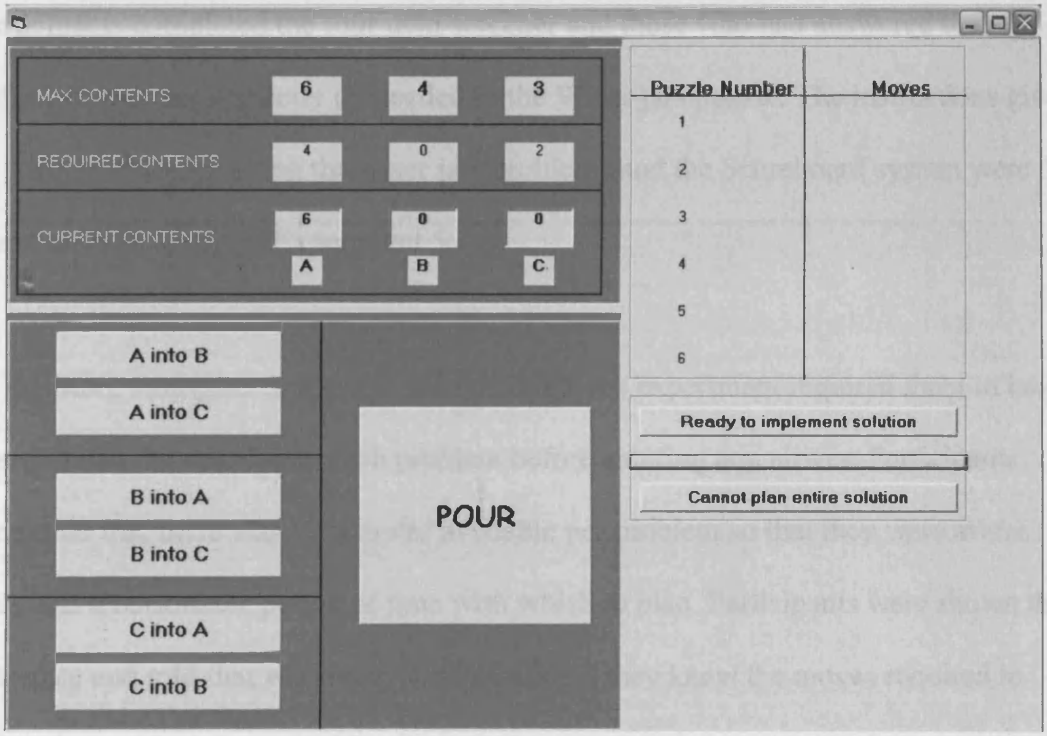


Figure 45. Screenshot of the Decision Interface for Total-Plan Users

The interface then required participants to make a simple choice. When certain that the solution and had planned in its entirety they could indicate this by pressing the button labelled 'Ready to implement solution'. If however, after a sustained period of planning participants were confident that all possible plans had been exhausted and were satisfied they could not generate the necessary plan, the button marked 'Cannot plan entire solution' was selected and problem solving began.

All three interface groups were informed that the aim was to complete the puzzle in as few "Moves", as shown by the scoreboard, as possible. A 12 minute time limit was set on each problem and a warning appeared on screen for a period of 6 seconds when 2 minutes remained in the trial.

Procedure

Participants completed the four item screener and those who had answered the sufficient number correctly proceeded to the Water jars puzzle. The instructions given to participants explaining the water jars problems and the Scoreboard system were the same as those given in Experiment 5.

In addition, total-plan users were informed that the experiment required them to have pre-planned the solution to each problem before entering any moves. Participants were told that there was 12 minutes available per problem so that they were aware they had a reasonable period of time with which to plan. Participants were shown the interface and told that when they were confident they knew the moves required to solve the problem they would indicate so by pressing the appropriate button and begin entering the moves. If during the implementation of moves they realised they had been incorrect or had made an error they should try and solve the problem in as few moves from that point onwards as possible. They were also informed that if they felt unable to generate a plan after a sustained pre-planning period, they would indicate this by pressing the appropriate button and could begin solving the problem. If this situation arose they were asked to solve the problem in as few "moves" as possible, as indicated by their score for that particular trial on the scoreboard.

Results

This section is divided into two main categories. As in Experiments 5 and 6 the main analyses are presented first. After the main analyses, there are also a number of supplementary measures that were recorded from total-plan users' behaviour. These include total pre-plan time, play time and the decisions on whether or not they could generate a plan for a particular problem as well as relationship with solution accuracy.

General Performance Data

Examining the number of successful completions of trials revealed that 81.37% of all trials were completed before the time limit expired. A breakdown by interface group revealed that the number of trials successfully completed were 80.88%, 85.29% and 77.94% for 1-Move, Look-ahead and Total-Plan interface users respectively. A Chi-Square test on number of solvers versus non-solvers by interface group revealed no significant difference ($\chi^2 = 1.23$, $p > .5$, $df = 2$).

Combining all groups' performance in terms of completed trials revealed 80.39%, 96.08%, 90.19% and 58.82% completion rates for trials 1 – 4 respectively, with trial 4 taking 7 seven moves to completion being the most difficult problem. A breakdown of completion rate by trial for each interface group can be seen in table 7 below.

Table 7.

Percentage of trials successfully completed for all three interface groups

	Trial			
	1	2	3	4
1-Move	88.23%	94.11%	82.35%	58.82%
Look-ahead	82.35%	100%	94.11%	64.70%
Total-Plan	70.58%	94.11%	94.11%	52.92%

The number of errors in proposed plans for Look-ahead users once again indicated the high accuracy of intended move sequences. Modal number of errors in plans by trial and by participant was zero. This measure will not be discussed further.

Optimal Performance Data

The measure of problem solving efficiency in terms of number of trials solved in the minimum number of moves indicated again that planning groups were much more efficient than 1-Move interface users. Out of a possible 68 trials attempted, only 12 were solved in the minimum number of moves by 1-Move participants. Planning groups solved almost equal numbers of trials in the minimum number of moves, with 31 trials being solved optimally by Look-ahead users and 29 by Total-Plan users.

A Chi-Square test revealed a highly significant difference in the number of problems solved in the minimum number of moves by interface group ($\chi^2 = 14.04$, $p < .01$, $df = 2$). A complex comparison Chi-Square, combining plan groups' performance and

comparing with 1-Move interface performance revealed a significant difference (χ^2 = 13.9, $p < .01$, $df = 1$).

Number of Excess Moves

Number of excess moves, m, was $\log_{10}(m + 1)$ transformed to reduce the violation of the normality assumption for ANOVA. The missing values that were due to unsolved problems were replaced with the grand means of the transformed values (cf. Delaney et al., 2004). A similar transformation and replacement process was also applied to missing total-time data. The impact of interface on number of excess moves made by interface groups is shown below in Figure 46.

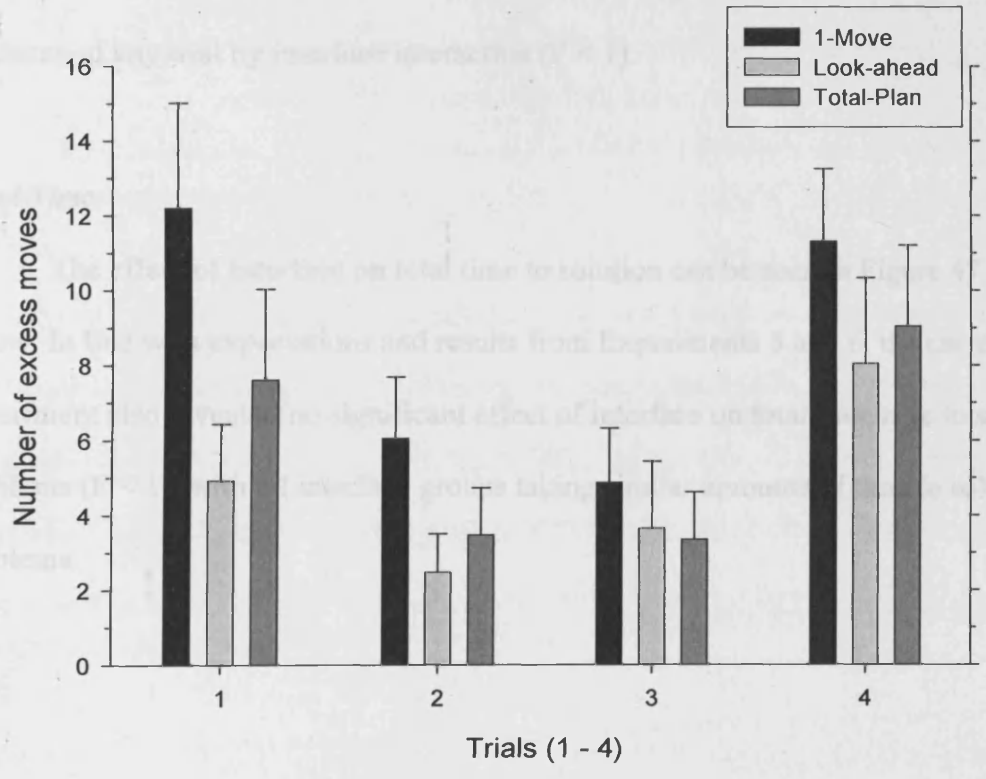


Figure 46. Number of Excess Moves over Trials by Interface Group

Error Bars Represent One Standard Error

A two-way repeated measures ANOVA revealed a significant effect of interface on number of excess moves, $F(2, 48) = 4.83$, $p < .02$, $MSE = 1.59$. Post hoc bonferroni corrected tests revealed a significant difference between look-ahead interface users and 1-Move interface users ($p < .02$). However the difference between Total-Plan interface and 1-Move controls was only marginally significant ($p = .057$). The analysis revealed no evidence of any differences between look-ahead and total-plan interface users.

As expected the analysis also revealed a significant effect of trial, $F(3, 144) = 4.31$, $p < .006$, $MSE = .76$. Post hoc tests revealed a significant difference between problem 2 and problem 4 ($p < .03$) and between problem 3 and 4 ($p < .02$). There was no evidence of any trial by interface interaction ($F < 1$).

Total Time

The effect of Interface on total time to solution can be seen in Figure 47 below. In line with expectations and results from Experiments 5 and 6, the current experiment also revealed no significant effect of interface on total time take to solve problems ($F < 1$), with all interface groups taking similar amounts of time to solve all problems.

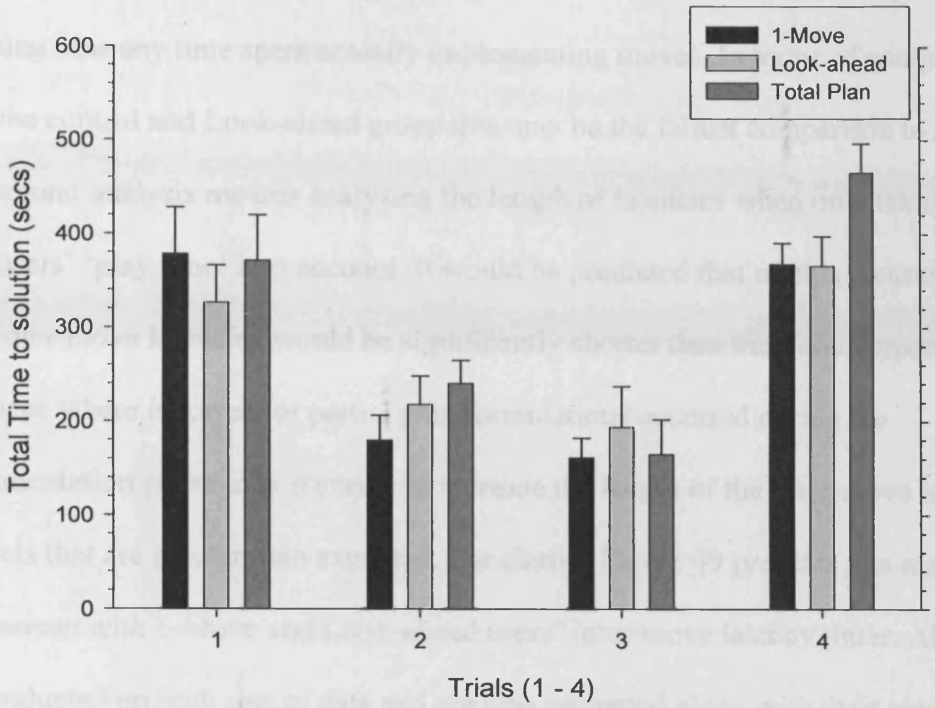


Figure 47. Total Time to Solution for the Three Interface Groups
Error Bars Represent Plus One Standard Error

N.B. For total-plan users the total time referred to here is calculated from their total time spent generating a plan plus their total 'play time'

The ANOVA revealed a significant effect of trial, $F(3, 144) = 20.79$, $p < .001$, $MSE = 1.49$. Pairwise comparisons on problem trial revealed trial 1 to be significantly different from trials 2 ($p < .003$) and 3 ($p < .001$) but not trial 4. Trial 2 did not differ from trial 3 but did differ from trial 4 ($p < .001$). Trial 3 also differed significantly from trial 4 ($p < .001$).There was no evidence of any trial by interface interactions ($F < 1$).

Inter-move Latency

There are a number of comparisons available regarding inter-move latency times. Given the process that Total-Plan users went through to reach a solution, two possible means by which examining their latency data exist. Firstly, as presented in

Figure 48, inter-move latencies can be calculated based on the total time spent pre-planning plus any time spent actually implementing moves. In terms of comparisons with the control and Look-ahead group this may be the fairest comparison to make. The second analysis regards analysing the length of latencies when only taking Total-Plan users' 'play time' into account. It would be predicted that on this measure Total-Plan inter-move latencies would be significantly shorter than their counterparts. Instances where incorrect or partial plan formulations occurred during the implementation phase may contrive to increase the length of the inter-move latencies to levels that are greater than expected. For clarity, Figure 49 presents this measure in comparison with 1-Move and Look-ahead users' inter-move latency times. ANOVA's are conducted on both sets of data and are also presented along with their respective figures. For analysis purposes the data were $\log_{10}$ transformed to stabilise the variance and missing trials were left with their original values due to a number of moves having been made by all groups on all trials therefore providing some measure of inter-move latency time, however infrequent they may have been.

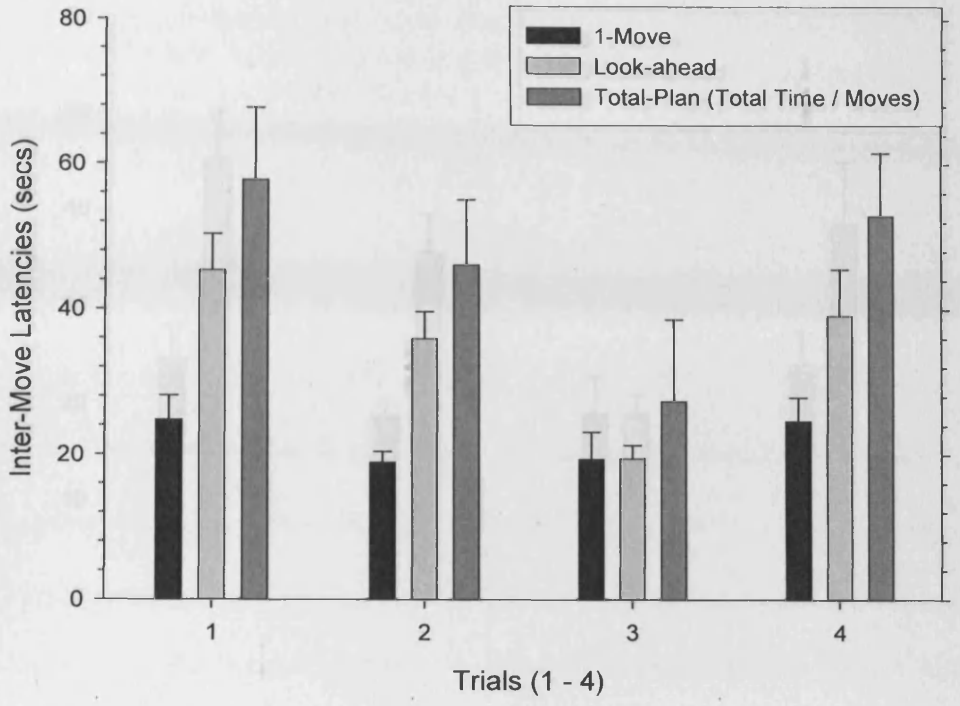


Figure 48. Inter-Move Latency Times for Interface Groups

Error Bars Represent Plus One Standard Error

As expected, the ANOVA revealed a highly significant effect of interface, $F(2, 48) = 9.39$, $p < .001$, $MSE = 1.12$. Bonferroni corrected post hoc tests revealed that 1-Move users had significantly shorter inter-move latencies than Look-ahead ($p < .005$) and Total-Plan users ($p < .001$). There were no significant differences between the two plan groups in the current comparison ($p > .1$).

There was a significant effect of trial, $F(3, 144) = 22.30$, $p < .001$, $MSE = 1.02$. Pairwise comparisons on trials revealed that only trial 3 significantly differed from all other trials (all p's $< .05$), with no other differences being significant.

The second analysis, this time comparing the inter-move latencies of Total-Plan (Play Time / Moves) latencies were calculated and are shown below in Figure 49.

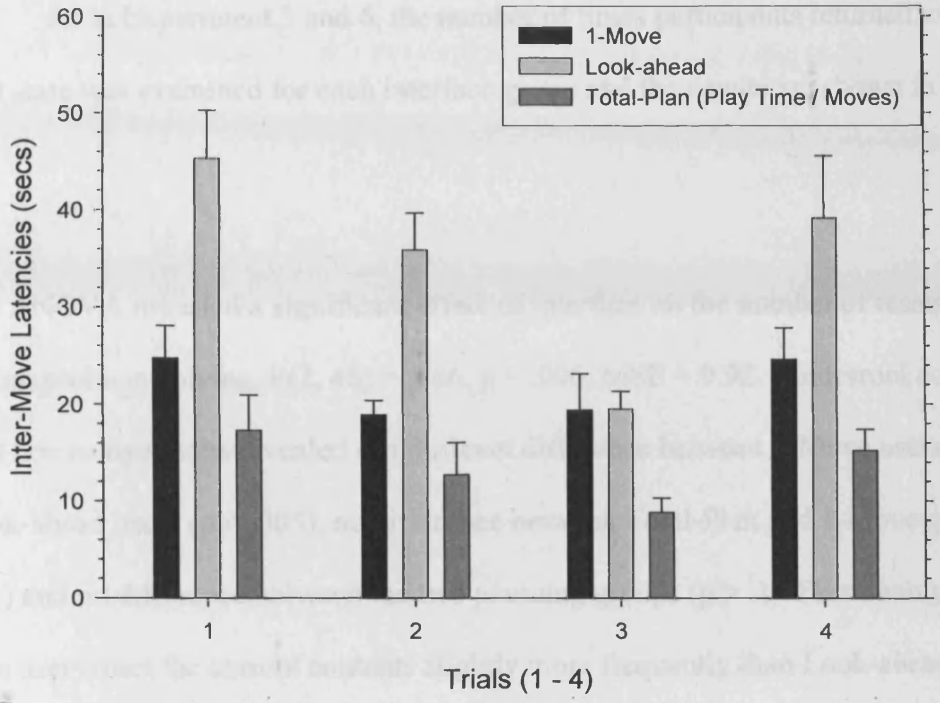


Figure 49. Secondary Comparison of Inter-Move Latencies with Total Plan (Play Time / Moves) Data

Error Bars Represent Plus One Standard Error

The ANOVA revealed a significant effect of interface, $F(2, 48) = 33.87$, $p < .001$, 3.77. Post-hoc tests revealed significant differences between Total-Plan and 1-Move users ($p < .001$), and also Look-ahead users ($p < .001$). The difference between 1-Move and Look-ahead users was also significant ($p < .003$).

The analysis revealed the effect of trial to be significant, $F(3, 144) = 14.17$, $p < .001$, MSE = .66. Pairwise comparisons revealed significant differences between trial 3 and all other trials (all p's < . 05), with no other significant differences.

Number of Resets

As in Experiment 5 and 6, the number of times participants returned to the start state was examined for each interface group and the results are shown in Figure 50.

The ANOVA revealed a significant effect of interface on the number of resets made during problem solving, $F(2, 48) = 5.66$, $p < .006$, $MSE = 9.92$. Bonferroni corrected post hoc comparisons revealed a significant difference between 1-Move users and Look-ahead users ($p < .005$), no difference between Total-Plan and 1-Move groups ($p < .1$) and no difference between the two planning groups ($p > .1$). Presumably, Total-Plan users reset the current contents slightly more frequently than Look-ahead users but not significantly so. This may happen when an error occurs when attempting to initiate a generated plan. Upon discovery of having implemented the plan incorrectly or realising an error in the proposed plan, Total-Plan users reset the current contents to begin the implementation phase anew. This would be less likely for Look-ahead users if making use of greater partial plans.

There was a significant effect of trial, $F(3, 144) = 5.84$, $p < .001$, $MSE = 5.05$, which was moderated by a significant trial by interface interaction, $F(6, 144) = 3.34$, $p < .004$, $MSE = 2.89$. Simple main effects analysis revealed a significant difference between the look-ahead and 1-Move users at trial 1, $F(2, 48) = 5.68$, $p < .01$, $MSE = 16.61$, but no differences with Total-Plan users. No other effects were significant (All p's $> .1$).

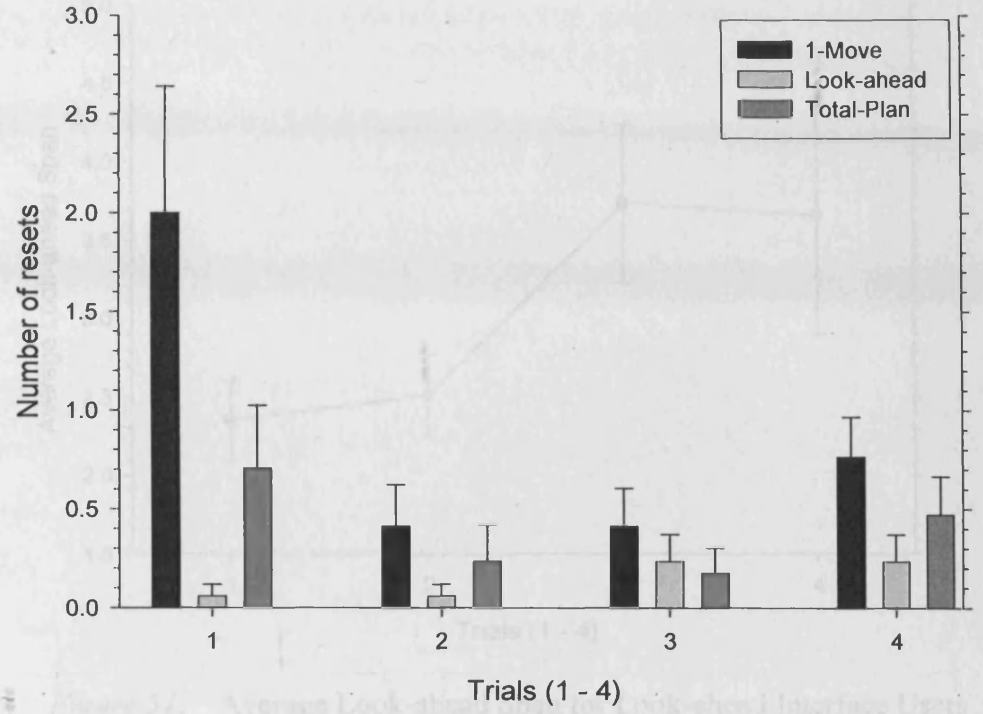


Figure 50. Number of Resets Over Trials by Interface Group

Error Bars Represent Plus One Standard Error

Look-ahead users on average therefore made less resets over trials than the 1-Move or total-plan users.

Look-ahead Span

Once again, the number of moves specified at once was examined for Look-ahead interface users and in line with the findings from Experiment 5 there was an increase over trials. They are also consistent with the previous current estimates of the average maximum number of moves searched along a solution path over trials. However, these results must be treated with caution due to the confound of minimum number of moves required to solve trials.

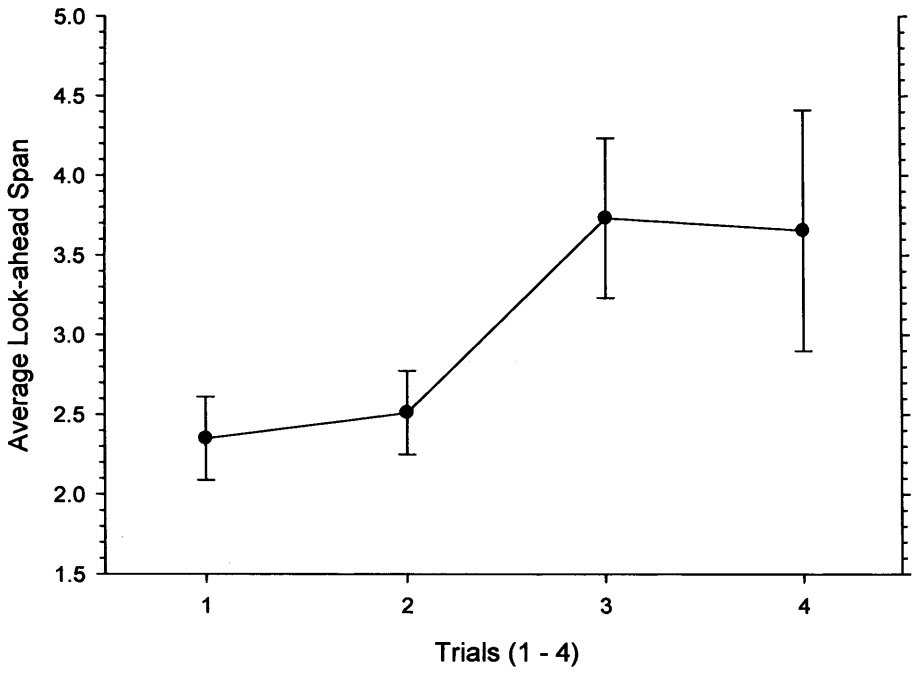


Figure 51. Average Look-ahead Span for Look-ahead Interface Users

Error Bars Represent Plus and Minus One Standard Error

A repeated measures ANOVA on look-ahead span over trials revealed a significant effect of trial, $F(3, 48) = 3.97$, $p < .03$, $MSE = 9.17$. The analysis also revealed a significant linear component to the curve, $F(1, 16) = 5.59$, $p < .05$, $MSE = 22.51$. There was no evidence of a quadratic component to the curve, $F < 1$.

For Look-ahead users, the proportion of completed trials were further analysed to examine the individual role of look-ahead (see Table 8). The proportion of participants completing the puzzle in number of pours was examined to further illustrate the typical behaviour being exhibited by look-ahead interface users during problem solving. The completion of the puzzle with only one pour press would again indicate that all moves were specified and entered at once whether using the minimal solution path or another almost equally viable alternative.

Table 8.

Proportion of trials solved in number of pours from all completed trials

	Trial			
	1	2	3	4
1 Pour	.285	.294	.312	.363
2 Pour	.142	.352	.25	0
3 Pours	.142	.058	.25	0
4 Pours	.142	.117	.062	.09
5 Pours	.071	.058	0	.09
6 Pours	0	0	0	.09
7 Pours	.071	0	0	0.9
> 7 Pours	.142	.117	.125	.272

The results are very consistent with the results from Experiment 5, with almost one third of all participants entering the necessary moves in one long sequence. The proportions using two or three moves to solution for trials 1 to 3 also indicate the breaking down of solutions that allow the completion of the problem from a state that may be deemed desirable due to its place on the solution path.

Supplementary Analysis for Total-Plan Users

Total-Plan Performance Measures

A number of supplementary measures from total-plan user's data were collected due to the slightly adapted methodology from that used by Delaney et al. (2004). Some of these measures involved recording the basic responses given by participants to indicate if they believed they had the required plan to their current

water jar problem. Table 9 below shows a breakdown of the 17 user responses for each trial and their relationship to actual problem solving performance.

Table 9.

Number of Can vs. Cannot Plan Responses, Relationship to Trials Completed and to Trials Solved in Minimum Number of Moves

	Trial			
	1	2	3	4
Indicated Had Plan	16	13	13	12
Indicated No Plan	1	4	4	5
Plan & Completed Trial	12	13	13	6
Plan + Solved in Min. Moves	8	7	9	4
No Plan & Solved	1	3	3	2
No Plan & Min. Moves	0	1	0	0

It therefore appears that an indication of a complete plan being specified does not necessarily equate to plan quality or accuracy. In fact, it almost appears that those who indicate they could not plan were more realistic than those who indicated they had an entire plan to a problem.

Decision Time for Total-Plan Group

Due to the distinct phases in terms of indicating when a plan had been formulated versus the amount of time spent implementing moves, further analysis can be carried out to examine differences in time spent planning per problem. Firstly, the time to decide whether a solution had been formulated versus the decision that it could not was examined and is shown below in Figure 52.

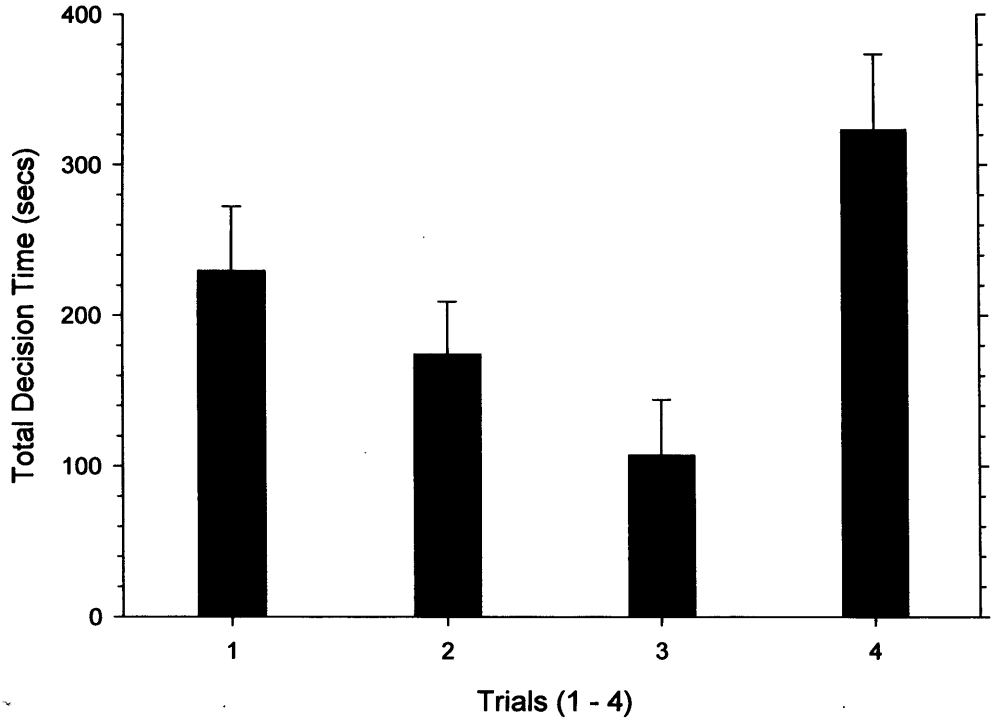


Figure 52. Total Time taken for Total-Plan Interface Users to Indicate if they Could Plan the Entire Solution or Not

Error Bars represent plus one standard error

A repeated measures ANOVA across trials of the $\log_{10}$ time taken to indicate if a plan had been formulated or was not possible revealed a significant effect of trial ($F(3.\ 48) = 5.00$, $p < .004$, $MSE = 1.08$). Pairwise comparisons across trials revealed trial 1 was no different from any of the trials, although it did approach significance with trial 3 ($p < .07$). Trial 2 was significantly different from only trial 3 ($p < .03$) and trial 3 was significantly different from trial 4 ($p < .02$).

Total 'Play Time' for Total-Plan Group

As previously shown, not all trials were completed and the graph shown below is the amount of time spent playing (whether action was taking place or not) after the initial decision was made by participants. Some of the play times extended right up until the time limit expired. Data were $\log_{10}$ transformed to stabilise for variance and a repeated measures ANOVA across trials was performed.

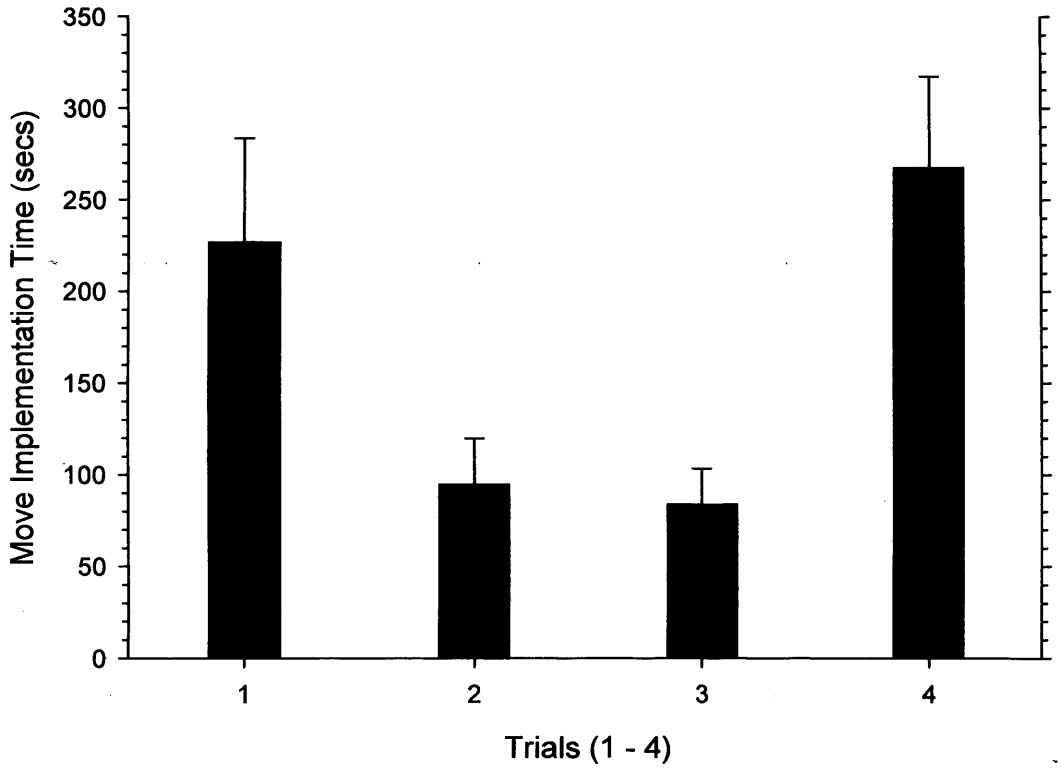


Figure 53. Total 'Play Time' for Total-Plan Interface Users

The analysis revealed a significant effect of problem trial, $F(3, 48) = 5.27$, $p < .003$, MSE = 1.07. Pairwise comparisons revealed that no difference in time spent implementing moves from any of the other trials. Trials 2 ($p < .008$)and 3 ($p < .004$) were significantly different from trial 4 and did not differ from each other.

Discussion

Experiment 7 confirmed several findings from the previous two experiments, along a number of important problem solving measures. At the most basic level, a significantly greater number of trials were solved in the minimum number of moves by plan based interface users. The slight adjustment to the interface from Experiment 5, in order to facilitate move implementation, appears to have had a more significant impact upon 1-Move users than both planning groups. Comparing the proportion of trials solved optimally from Experiment 5 for 1-Move users revealed a drop from .33 to .17. Look-ahead performance however remained unaffected by the change in interface (.46 for both experiments). The total-plan users' performance closely followed at .42. These results suggest that the argument from Experiment 5, regarding cost of implementation being at least one factor for increased 1-Move performance, appear to have at least some support from the current findings.

The results also prove interesting when comparing the performance of both planning groups. An almost identical number of trials were solved in the minimum number of moves by both planning groups, with both groups being significantly more efficient in their performance than 1-Move control subjects. The result appears to indicate given that no differences were found between planning groups, those in the Total-Plan group were no more efficient in the generation of accurate plans than users who may only partially plan their solutions.

The lack of a clear significant finding in relation to excess moves for Total-Plan users is most likely due to a number of competing factors. The first possible factor could be in relation to the smaller number of participants being used per group

compared to previous experiments. There were also fewer numbers of test trials in the current experiment than previously used. Combined with the efficient problem solving performance of all groups at trial 3 would suggest that either increasing the number of subjects taking part in each of the interface groups or extending the number of trials to be completed would have indeed revealed a significant difference between both planning groups and controls, rather than look-ahead subjects only.

Alternatively, the lack of an outright significant difference in the measurement of excess moves between total-plan users and 1-Move users may in some part be due to differences the Scoreboard manipulation has upon 1-Move performance. Comparing the control subjects as used by Delaney et al. with those used in the current experiment, current 1-Move users may at least have a partial motivation for performing at a higher level than they would do normally. The lack of visual feedback, as in the design from Experiment 6, indicated that when participants are free to enter moves as they wish without any cue as to how they are performing, problem performance worsens. The control subjects used in the Delaney et al. (2004) study were simply told to solve the puzzle as best they could, similar to control subjects from experiment 6. Comparing the Delaney et al. instructions for their control subjects with their total-plan users who were not allowed to begin implementing a single move until all moves to a solution had been pre-planned, may have contributed to the large performance differences reported in their study. It may be possible when compared with a control group who are perhaps a much fairer comparison, may reduce the effect and lead to smaller differences in the number of excess moves made. When given no inclination or mechanism to increase performance beyond any basic minimum and a lack of visual feedback that may on

some levels promoted excessive move making, controls in the Delaney et al study would have most likely just had the aim of finishing the puzzle using any method possible and as quickly as possible. The comparability of controls in the current study would appear to be a much greater test of the manipulation to generate total plans when the control group is also given some incentive to perform with some degree of effort.

The most likely reason for a significant difference of the current look-ahead manipulation would therefore appear to lie not only in the number of optimal trials solved but rather in increasing the performance of participants in trials that were not solved in the minimum. Giving participants the opportunity to adapt to each problem as they see fit and to excel whenever possible appears to reduce the numbers of excess moves made. The Total-Plan manipulation would not support participants who are unable to generate the entire solution path, but who may benefit more from the opportunity to interleave planning and acting. This approach, which requires less intensive mental simulation, allows performance to be dictated not by the interface but by the decision of the current user and skill of the participant on a case by case basis. Therefore even though participants may show an indication of increasing look-ahead over trials they can also choose to reduce planning to more local based instances if they cannot see a global solution path.

The predicted difficulty for total-plan users in terms of finishing puzzles with longer solution paths was only partially supported in that fewer numbers completed trial 4, requiring 7 moves to solution, although all interface groups appear to be significantly affected by trial 4 in particular. The larger number of non-solvers in trial 1 performance for the total-plan interface also suggests that initially participants or a

subset were finding it difficult to adapt to the task. This result is reminiscent of the results from trial 1 performance of look-ahead users from the 8-puzzle in experiment one. The requirement to look-ahead by 3 steps was a difficult task but one that participants soon adapted to by trial 2. The results and indeed slight performance increase by trial 2 suggest that planning and look-ahead is a skill that may develop over time, even after only a very short period of exposure to a task.

Total solution times were largely unaffected by manipulations to increase planning. There was some evidence that trial 7 for total-plan users was more difficult than it was for 1-Move and Look-ahead users. However, all groups appeared to have particular difficulty in reaching solutions for this problem. This finding is unsurprising given the previous findings for solution time. This measures is undoubtedly unaffected due primarily to the problem characteristics of the water jars.

Of interest concerning the plan and implementation times of Total-Plan users is the relative inefficiency demonstrated in their 'play' performance. Trial 1 implementation time performance suggested that only partial, inaccurate or incomplete plans had been formulated by Total-Plan users. The move implementation phase appears to have been used as a period of rehearsal for intended plans or to generate new plans, rather than demonstrate the fast uninterrupted behaviour that final path behaviour usually typifies (Kotovsky et al., 1985). This may have been due to simple memory failure of intended plans or due to the realisation after entering a first move that the next sequence of planned moves were in fact inaccurate. Such a realisation would require a new plan to be reformulated from the beginning and may explain the larger latency times than would have perhaps been expected. Such an observation on the reliance of

the external environment to provide backup to inaccurate plan information is consistent with evidence from rational analysis and current findings regarding the trade-offs between accurate information contained within the environment and that contained "in-the-head" (Gray & Fu, 2001). While plan formation for Look-ahead users has consistently been found to be very accurate these plans have been more partial in nature and therefore the perfect knowledge "in-the-head" is enough to guarantee accurate problem success.

Chapter 5

General Discussion

Introduction

The purpose of studying the role of look-ahead during problem solving centred around two motives in particular. The first was to find direct evidence to support recent arguments that increased planning and look-ahead were responsible for observations of improved problem solving performance. Additionally, the research attempted to develop interface based manipulations that would bring about the previously observed problem solving behaviours. The second motivation, presuming the first aim could be demonstrated, was to provide data about the characteristics of the look-ahead process. Identifying features such as its general functioning during problem solving, average look-ahead span, changes in average span over time or with increasing experience and relationship to problem performance were all key points of study.

Summary

In an initial series of four experiments the 8-puzzle (Ericsson, 1974), was used as a preliminary task to explore some of the initial research questions. Experiment 1 explored the simple presumption that if increased look-ahead was at least partly responsible for greater problem solving performance (e.g. Delaney et al., 2004; O'Hara & Payne, 1998), extending the amount of look-ahead required from participants would automatically lead to positive benefits for performance. However, the results ran counter to this hypothesis and revealed no overall benefits in terms of fewer moves to solution compared with control performance. While inter-move

latency times revealed that the planning manipulation appeared to be having the desired impact in terms of increasing the amount of time spent planning moves, no performance benefits appeared to result as a consequence of the increased planning time. For look-ahead users, inter-move latency times were still relatively large by the end of trials, suggesting that a quick adaptation to the task was not taking place, although total time to solution was largely unaffected by the increased latencies. The conclusion from this initial study suggested that any attempts to enforce a look-ahead span may not work due to the complete freedom in terms of flexibility needed for plan formulation. A strict enforced span length may never work and instead a discrete manipulation that allowed for individual decision making regarding time spent planning and acting was then developed.

Experiment 2 therefore sought an alternative approach from that undertaken in Experiment 1. The rather laboured performance of look-ahead users on the previous trial 1 performance suggested that the initial demand to specify three moves was perhaps too taxing for many participants. A second possibility was that even if participants could enter three moves at a time they could not use such sequences with such a constant length to generate more efficient plans or solutions. A more likely argument is that increased look-ahead would only occur at the choice of the participant and perhaps only at specific times when judged to be of most benefit. Attempting to dictate behaviour in an area such as planning that has been shown to be best categorised by opportunistic and flexible thinking may only lead to difficulty for participants.

A 'Scoreboard' system was introduced in Experiment 2, in an attempt to increase the desire of participants to plan and possibly tap into the more likely

approach of planning opportunistically (Hayes-Roth & Hayes-Roth, 1979). The use of the Scoreboard system to reward greater depth of look-ahead and planning with lower scores per trial resulted in the desired effect of improved problem performance, with large increases in look-ahead depth observed by the end of trials. The results showed that lengthy sequences of tiles could be specified by participants, far above the three steps from Experiment 1, but suggested that look-ahead depth is best decided upon by the respective participant at any given time. However, a subset of participants were unable to complete all trials and so Experiment 3 introduced the additional elements of a time limit to allow exiting from a trial that cannot be completed and a hint which aimed to increase the conceptual knowledge of participants and reduce the possible reliance on unproductive or weak strategies. It also examined another factor, that of solution path length, as a possible reason for previous participants' inability to solve the 8-puzzle. Differences in the number of short versus long solution path problems solved would indicate that solution path length may have been a contributing factor in previous failed solution attempts.

Experiment 3 found no difference in the number of trials solved with short or long length solution paths, indicating solution path length was not the main factor for problem difficulty encountered by subjects in Experiment 2. However, a number of trials were not solved within the time limit suggesting that participants would still need a means through which to exit a problem should a solution not be forthcoming within a reasonable period of time. The final 8-puzzle experiment examined the development of look-ahead over a greater number of trials, all requiring 17 moves to solution. The results suggested that look-ahead develops over trials irrespective of changing start state suggesting more general heuristics were being used to increase

performance rather than the specific mechanisms of memory and chunking that were most likely responsible for the large look-ahead span measures taken from Experiment 2. On trials where new start states are presented per trial there appears to be a limit of between 3 and 4 moves for the depth of discretionary look-ahead search which is in agreement with the existing conceptualisation of look-ahead as a limited process that adapts and reacts to and with changes in the environment and to the development of knowledge in a task with increasing experience (Hayes & Simon, 1974).

While the evidence for any learning over trials is limited both due to the range of experiments using the 8-puzzle and the necessity to implement a hint in an attempt to reduce non-completion rates, there are elements of the performance that have indications of learning. From Experiment 2, the rapid decrease in inter-move latency times for look-ahead users suggest that a more procedural phase of learning was being entered into with singular moves being combined into longer productions in line with theories of expertise, learning and cognitive skill acquisition (e.g. Anderson, 1990; VanLehn, 1996, 1989). The evidence from Experiments 3 and 4 also show greater use of larger sequences of tiles, again suggesting the compounding of particular moves with general problem solving heuristics being employed to increase performance while also perhaps reducing working memory demands of the task.

Average versus Maximal Look-ahead

While average look-ahead has been examined in earlier chapters, the characteristic of 'maximal look-ahead' has not yet been discussed. Examining 8-puzzle performance, with particular attention to Experiments 2 and 4, a number of potentially important attributes of look-ahead maxima are revealed. The two most important considerations

centre around maximal look-ahead depth and individual differences for maximal search and their subsequent effects on performance.

The design of Experiment 2, whereby participants transformed the same start-state to a particular goal state over ten trials led to frequent observations of look-ahead spans much larger than typically considered possible during problem solving. In fact the depth of look-ahead search for some participants was of such a length that it contradicts the generally limited values that are assumed possible and as characterised during the current work. The data collected from participants HN, LS, FT, JC contain multiple examples of trials being solved using the shortest solution path but also by specifying all the necessary moves (17) at once, thereby achieving the lowest score possible on a trial of one. What is even more surprising is that this level of performance was becoming evident as early as trial 5 for some participants. It would appear that participants may have realised the efficiency of the 17-move solution path and did not spend any further time searching for an alternative that would be in accordance with a rational analysis explanation of performance (Anderson, 1990). As the solution path could not be improved upon the only elements that could be improved upon were in relation to trial score if a score of one had not yet been achieved, inter-move latencies and total time, all of which did decrease rapidly over the remaining trials. The observed performance increases are consistent with accounts of expert performance starting in a declarative state as plans take effort and explicit development to form before gradually transitioning to a more procedural state where individual steps are compiled into sequences that are automatically initiated without the need for deep thought or analysis indicative of expertise (Anderson, 1993). The iterative depth of search only, rather than breadth and depth in parallel, may also allow such lengths of look-ahead to be possible by reducing the working memory load

on participants (Delaney et al., 2004). The 8-puzzle also has a low underlying level of complexity in terms of the small impact of having to decide upon which moves are possible. Chaining sequences that are all naturally linked together when rotating a number of tiles may also allow for such large observed performance levels.

As has become a consistent finding during the current work, the need for flexible interfaces appears to be a necessary feature to make the most of planning. The evidence from Experiment 2 in relation to look-ahead maxima highlights the importance of this feature. However, while the above participants all exceeded typical look-ahead values, a number also demonstrated very low look-ahead depths throughout the 10 trials but with very different overall performance levels. Participant LW did not exceed specifying 7 moves on any of the ten trials. In fact the modal depth of look-ahead on the final trials, in terms of number of tiles physically specified, remained at a limited depth of only 1. However, performance was highly efficient with the puzzle often being solved in 21 moves, only 4 over the minimum possible of 17. The participant JG also demonstrated a limited maximal look-ahead span with a modal look-ahead maximum of one step. However, the performance of JG was substantially less efficient than that demonstrated by LW, with final trial performance taking 67 moves. This would indicate that other, possibly more general problem solving mechanisms, were also in operation for the participant LW that still allowed near optimal performance whilst participant JG may not have discovered any additional strategies by which to improve performance.

Evidence for such an argument can be found in the data of both Experiment 2 and Experiment 4, indicating that although maximal look-ahead span had reached a certain plateau near the end of trials for many participants, there is clear evidence from the move data that performance was continuing to improve. This finding

suggests that an alternative mechanism(s), in conjunction with depth of look-ahead search, was operating during the course of the trials. This would seem not only highly probable but comparable to the earlier description of expert chess players using not only look-ahead but more general mechanisms known as general 'templates' that aid performance by recognising and grouping promising states or possibly in this case sequences of tiles that aid minimise the demands placed on memory (Gobet & Simon, 1996). Such recognition of useful template configurations may also explain why participants with much more limited look-ahead span could still exhibit strong problem solving performance while keeping look-ahead depth relatively low.

The examples given above have suggested the larger the value of look-ahead the better the problem solving performance. There was also evidence that being able to specify large numbers of tiles does not necessarily lead to the most efficient performance. The participant CGG demonstrated a maximal look-ahead span of 27 moves on the final three trials. From trial 6 onwards, CGG began solving the puzzle in 27 moves. On trials 6 and 7, CGG had begun entering three distinct blocks of moves, giving a score of 3, before solving trials 8 to 10 by entering the entire block of 27 moves at once. An implication of such a finding is that if an interface encourages the memorisation of large chunks of moves to the point that it has become proceduralised and participants are content the solution path is optimal or near optimal then participant satisficing may occur, when further planning and look-ahead would be of more benefit. While such an extreme example was generally uncommon other look-ahead participants' solution paths demonstrated these characteristics, albeit entering moves in lesser sized chunks than that of CGG. For example, participant JGG generally solved the final five trials using around 30 moves per trial, using

various solution paths per trial. Look-ahead depth was consistently around 7 or 8 moves on average per implementation suggesting that a certain fixation or preference for particular blocks or sequences of tiles had firmly settled into plan development. Searching for a more efficient solution path had no longer become a priority as the current method guaranteed problem solving success and also with a very low problem solving Score.

The maximal look-ahead data for Experiment 4, which required the transformation of 8 different start states to a goal state, followed the patterns of those described from Experiment 2. There still remain large individual differences in relation to maximal look-ahead while the depth for some participants remains unusually large albeit not to the same extent as exhibited by a significant proportion of participants from Experiment 2.

The conclusions from the first series of experiments therefore were that dictating a range of steps to be searched does not necessarily lead to improved performance. The results indicate only a limited range of steps are initially considered, although the depth of search can rapidly increase from that point onwards when dictated by participants. With the appropriate motivation in place, in terms of encouraging greater planning, the typical depth of look-ahead will increase linearly over trials when problem start and goal states remain static. A more reasonable estimate of average look-ahead, when problem start states are in a state of constant change, is more likely to reach depths of approximately between three to four steps on average.

In three further experiments in the domain of solving 'Water Jars' problems, several findings from the 8–puzzle experiments were confirmed, but in a new problem

solving domain. Experiment 5, using the scoreboard system from the previous experiments to encourage look-ahead, was successful in increasing problem solving performance as evinced by a reduction in the numbers of excess moves required to solve puzzles. In addition to this, look-ahead interface users solved greater numbers of problems in the minimum number of moves while leaving both completion rates and total time to solution unaffected when compared to control performance. Large look-ahead spans for a proportion of participants were also accompanied by evidence not only of efficient performance but also of largely error free performance.

Experiment 6 departed from the use of additional aids to increase performance, such as the Scoreboard system, with a mechanism tied directly into the interaction between participant and problem solving interface, commonly referred to as System Response Time (SRT). In line with previously reported findings it was hypothesised that delaying the system response time to user input would have an effect on the depth of look-ahead search that participants would perform. The manipulation was successful in improving performance compared with controls but not to the extent as previously observed. Look-ahead values were also inconsistent with the performance of participants from Experiment 5. Results were interpreted as indicating that the short SRT of only 4 seconds lead to an active adaptation to the task by participants. Participants actively decided not to enter or perhaps even plan to the extent indicated previously as the perceived utility of such an action was judged to be of less benefit. It appears that although performance was still significantly greater than controls the depth of look-ahead was more moderate in comparison to that previously demonstrated. Previous manipulations of SRT (e.g. Child, 1998), to that used in Experiment 6, have found greater effects upon performance and the current argument would hypothesise that increasing SRT during water jar tasks would also

lead to participants adapting levels of planning and look-ahead to those found in Experiment 5.

In Experiment 7, the final experiment of the thesis, three groups' performance was directly tested. As used in the scoreboard design of Experiment 5, the performance of a control and look-ahead group were again compared. A second plan group (total-plan) was also added, with participants being required to pre-plan the entire solution, if able, before move implementation could commence. Whilst all subjects used the scoreboard system, the instructions to plan were varied amongst the three experimental groups. The results were all in the predicted direction in regards to planning groups' performance compared to controls, with greater numbers of trials being solved in fewer excess moves, although total-plan performance was only marginally significant. A greater number of trials being solved in the minimum number of moves possible was also observed for both plan groups, again in line with predictions. Total time to solution was unaffected for planning groups when compared to 1-Move control subjects. There was only limited evidence, all of which was non-significant, when comparing the performance between look-ahead and total-plan users. The prediction that total-plan participants' performance would decrease with increasing solution path length was largely unsupported. Completion rates over trials were comparable between control subjects and look-ahead users. All three experimental groups appeared to be affected equally by the difficulty of trial 4 that required 7 moves to solution. All groups required greater numbers of excess moves on this trial with a large number of participants in all conditions failing to complete the trial within the time limit. While overall number of excess moves was only marginally

significant for total-plan users, this may be due to the small number of experimental trials and fewer participants per group, so firm conclusions cannot be made.

Look-ahead values for Experiments 5 and 7 showed remarkable consistency with the data collected from Experiments 2 and 4. Look-ahead span showed a similar pattern of development with a depth of fewer than three steps being considered but increasing linearly with increasing levels of task experience. The interpretation of look-ahead data for Water Jars performance is complicated by the differing numbers of moves required to solve problems. However, a limit of approximately four steps was observed by the end of trials on Experiments 5 to 7, supporting previous data from the 8-puzzle experiments and suggesting the average limit of look-ahead span in novel domains with restricted experience. However, as reported in the results section for the water jars experiments incorporating the scoreboard manipulation a large individual variability within the typical depth of look-ahead search is evident.

Problem Solving Success and Problem Characteristics

The success of the current manipulations to increase planning, look-ahead and improve performance stand in stark contrast to the studies discussed in the introduction in relation to the adapted TOL task developed by Ward & Allport (1997; TOL-R). Several reported studies demonstrated no evidence of improved performance in relation to problem performance with increased time spent planning. The question remains as to why such a marked difference in performance between look-ahead participants versus control subjects in the current work exist when no such differences have been found between participants on a task that is primarily used as a test of planning.

As described in the introduction, there appears to be evidence that the TOL-R problem actually has by its nature the ability to allow spatial rehearsal of moves that compensates for poor quality plans that may be very inaccurate or only partially formulated (Gilhooly et al, 1999). In a follow up study Gilhooly, Wynn, Phillips, Logie & Della Sala (2002) analysed individual performance differences on the TOL-R task and found that participants' performance on dual tasks that have a high visuo-spatial component correlated most significantly with a subjects subsequent TOL-R performance. Those performing with the highest visuo-spatial capacity in tasks such as the Corsi Block task performed most accurately on TOL-R performance. There was no such relationship between verbal tasks. Gilhooly et al. (2002) claim that the evidence collected from their studies taps into one of the subcomponents of the visuo-spatial system which they call the 'inner-scribe'. The goal selection strategy that they argue is responsible for TOL-R performance is "executed using a code to represent planned sequences of movements of discs, and that these planned sequences are held in the active spatial rehearsal mechanism (inner scribe) of visuo-spatial working memory" (Gilhooly et al., 2002, pp. 176). The longer latency times required to move discs has also been found by Phillips, Gilhooly, Logie, Della Sala & Wynn (2003) between young and old participants. The argument is that older participants who have more difficulty formulating accurate or plans of a greater depth are able to compensate for these weaknesses by the spatial mechanisms inherent in the problem task. Phillips et al. (2003) found no performance differences in terms of the numbers of problems solved in the optimal number of moves. The results also strengthen the argument that the TOL-R problem has been altered by the equalising of peg size (Unterrainer et al, 2005). The lack of a spatial correlation during testing on the original TOL task and tests designed to test spatial memory also suggest the link

between the new characteristics of the TOL-R task and the possible spatial mechanisms that result from the change to the problem space that equalising peg sizes may have had. The lack of any such available spatial rehearsal mechanisms in the current tasks would indicate one plausible explanation as to why performance differences appear to be more easily evinced. Without structural or perceptual cues as to the best next move, performance will largely be dependent upon plans of genuine quality or greater depth. What therefore are the implications for the TOL-R as a means by which to test planning? While the task may still be useful as a test of planning it appears that the implementation phase is responsible for perhaps masking the very data that it aims to reveal such as the inaccuracies contained within a participant's pre-plan. It may instead indicate that assessments of participants' levels of planning, in terms of depth or quality, are at inflated levels than they can actually accomplish. Therefore, the means by which the testing phase is implemented may need reconsideration as a means of assessing levels of planning. However, the nature of the pre-plan stage may also be compromised by spatial rehearsal mechanisms inherent in the task, yet results from verbal protocols suggest this is unlikely and any problems in plan formulation can still be detected.

Spatial rehearsal mechanisms are of course not the only means by which performance can be supplemented by task characteristics. Research conducted during the course of the current thesis but not reported here, used a problem known as the Car Park puzzle (for details see Jones, 2003). Results from the verbal protocols collected during problem solving indicated that large amounts of planning may naturally be induced by the problem itself through the problem solving strategy afforded by the problems structure and that a planning manipulation may add little advantage over control subjects. The applicability of interface manipulations to

increase planning are therefore not envisaged to either always be conducive to increasing performance or be applicable to all tasks.

Future Work

Empirical extensions to the current work can be divided into two distinct directions. The first provides suggestions for possible refinements and expansion to some of the existing experiments reported in the current work. The second direction explores future experiments that fundamentally change either the approach used in terms of the manipulations or use of new methodologies that may provide finer grained data to supplement the current problem based approach.

1. Refinements of current work

There are a number of refinements to the current work that would have been undertaken during the course of the thesis if time had permitted. One particular finding that still remains largely ambiguous is in regards to the performance of participants operating under the SRT manipulation from Experiment 6. While the performance benefits were still evident, overall problem efficiency was noticeably different from that observed in Experiments 5 and 7. As previously discussed the results could be taken as evidence that greater performance can still be influenced by increasing the time spent reflecting upon a proposed next move. Given the results of the other two water jars experiments and the 8-puzzle results as a whole this finding would not seem to be well supported. It is much more likely that the manipulation simply did not impact strongly enough upon participants to enter into deep problem space search or either enter a sequence of pre-planned moves of any length. Future work would therefore attempt to examine the adaptation of performance with

increasing lengths of SRT. It would be predicted that an active adaptation to the task in terms of greater performance would be observed with increasing lengths of SRT, while a natural limit for the utility of longer response time would also be predicted.

For water jars problems, a number of particular avenues for future work exist. The list system used by look-ahead users to specify their solution path may be considered in a weak sense to be an additional support for problem solving. Cary & Carlson (1999) for example reported that external supports for problem solving can often affect both performance and choice of problem solving strategy. The information contained while solving the water jars in the current experiments is only in symbolic form and cannot provide any immediate information regarding the current problem state or the exact values of any previously visited 'current states'. It may however exert some influence upon problem solving behaviour through the distribution of working memory resources that would otherwise be engaged if the itemised sequence of moves was not present (Cary & Carlson, 2001). A number of simple future manipulations would allow the examination of these issues. One possible manipulation would be to simply remove the list, yet still allow the specification of solution paths of any length. The ability to recover from a mistake by adapting the move sequence would be removed from such a manipulation and may actually encourage the active frequent rehearsal of solution paths which would be unwanted, so caution is warranted when making any such changes.

A third refinement would aim to increase the recording of data at the individual move level, particularly in regards to water jars problems. As a follow on from the previous point, the recording of data from look-ahead users' manipulations with any moves

contained in the list during a problem trial may prove worthwhile. Moves may have been stored, deleted and manipulated during the formulation of solution paths. Although little evidence of such behaviour was observed it would help either inform or rule out the use of the list box as an additional problem solving aid as opposed to being used specifically for move specification purposes. The average inter-move latency times were recorded during solutions but the capturing of data at the finer level of the individual move may reveal more specific patterns of behaviour and allow greater insight into the likely problem solving strategies being employed by participants. All these issues could be explored in any future experiments to explore look-ahead processes in the current tasks.

2. New Approaches

While the Scoreboard manipulation proved very effective in the current experiments it is limited in several respects. In particular, it is only as effective as the perceived strength or weight given to it by participants. There was evidence from a number of participants' performance measures and attitudes to tasks that they were not particularly encouraged to plan by the scoreboard system. While it appears to be generally effective, a greater more globally effective mechanism could be developed with an implementation that affects all participants more equally yet still allows for the necessity of flexibility and maximal adaptation. One such mechanism could be a keystroke limit, with the setting of the exact limit determining the flexibility while imposing a constant meaningful motivation to plan for all participants (e.g. Trudel & Payne, 1995).

A second future extension would be the use of verbal protocols as frequently used to study problem solving behaviour (e.g. Ericsson & Simon, 1991; Trudel &

Payne, 1995; VanLehn, 1991; Anzai & Simon, 1974). Asking participants to verbalise while problem solving could provide greater understanding of how look-ahead processes develop over time. The details of exactly when the look-ahead processes occur and develop over the course of a task(s) may be much more easily identified with such a methodology and provide a much finer level of detail and analysis.

While the work focused upon the more limited area of problem solving there are possibly a number of implications from the current research that may be informative for the more 'knowledge rich' domains that people more commonly use or interact with. As previously stated the nature of the environment will most likely play a large factor in deciding the applicability of certain design decisions. The current work would be difficult to implement in any current website that had a strictly commercial use. For example, a common rule of thumb for webpage design is the often cited: "Don't make me think" (Krug, 2000). The principle is to provide the user with a series of unambiguous choices that require little processing or thinking time, even if presented with a number of options. The current work would most likely find greater favour in an area that specifically required its users or would benefit from its users in fact thinking and planning more. A positive finding from the current work is that this extra effort does not necessarily need to come at the expense of time. The results presented in the previous chapters have consistently demonstrated that total time to complete a task does not have to be sacrificed in order to find evidence of a benefit for performance. Further still, the total time factor was unaffected at every level of experience with a problem, from the very beginning of trial one to the final performance on a final trial. Growing areas such as E-learning and distance learning may be one particular example that may benefit from manipulations such as those

used in the current experiments. Exploring the means by which learning materials and examples are presented to people with different levels of ability may be one avenue worthy of future study. The current mechanisms may also be used in tandem with newer approaches to learning. Recently for example, 'AVOW' diagrams have been developed (e.g. Cheng 1996; 1999) as an alternative means with which to represent and teach physics concepts and principles. Such learning materials are currently only taught using traditional paper-pencil tests. The use of interactive mechanism may enhance both the teaching and learning of such materials to improve performance further.

Conclusion

The decision to plan or not is not a binary choice. In fact, the decision or not to plan may not even lie with the participant. It may appear a simple conclusion but problem solving tasks that may benefit the most from manipulations designed to specifically increase planning and subsequent performance will be those that do not naturally induce planning through task features or characteristics but that would benefit almost immediately from it. The two tasks studied in depth in the current work both lack a subgoal structure which may provide an indication of the types of puzzles that may profit most from attempts to increase look-ahead. The implications of the current findings also suggest that naturally inherent mechanisms in a task should be examined in detail before the task itself is put forward as a means by which to examine human performance or cognition. The TOL problem is a prime example of one such task and its suitability as a true planning task appears to be called into question given the literature and the findings of the current work. The fundamental rule that may govern any successful interface manipulation may be that it will only be as successful to the

degree in which it supports both human opportunistic planning mechanisms and the willingness of a person to push their behaviour beyond the minimum. This however will only occur if fitting reason for such an increase is perceived by the participant and the benefits are made apparent.

References

Anderson, J. R. (1983). *The Architecture of Cognition.* Cambridge MA: Harvard University Press.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1993). *Rules of the Mind.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1993). Problem Solving and Learning. *American Psychologist, 48(1),* 35 – 44.

Anzai, Y. and Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86,* 124–140.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K.W. & Spence, J.T. (Eds.), *The Psychology of Learning and Motivation,* New York: Academic Press.

Atwood, M. E., Masson, M. E. J., & Polson, P. G. (1980). Further explorations with a process model for water jug problems. *Memory & Cognition, 8(2),* 182 – 192.

Atwood, M. E., & Polson, P. G. (1976). A process model for water jug problems. *Cognitive Psychology, 8,* 191 – 216.

Baddeley, A. D. (1986). *Working memory.* Oxford: Clarendon Press.

Baddeley, A. D. (1990). *Human memory: Theory and practice.* London: LEA Publishers.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In Bower, G.H. (Ed.), *Recent Advances in Learning and Motivation (Volume 8),* New York: Academic Press.

Berg, W. K., & Byrd, D. L. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and*

Experimental Neuropsychology, 24, 586-604.

Berry, D. E., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalized knowledge. *The Quarterly Journal of Experimental Psychology, 36A*, 209-231.

Berry, D. C., & Broadbent, D. E. (1987). The combination of explicit and implicit learning processes in task control. *Psychological Research, 49*, 7-15.

Bratko, I. (2000). *PROLOG Programming for Artificial Intelligence* (3rd Ed.). Longman Publishing.

Broadbent, D. E., FitzGerald, P., & Broadbent, M. H. P (1986). Implicit and Explicit Knowledge in the Control of Complex Systems. *British Journal of Psychology*, *77*, 33 – 50.

Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *The Quarterly Journal of Experimental Psychology, 55A*(1), 241 – 261.

Butler, T. W. (1983). Computer response time and user performance. *ACM CHI '83 Proceedings: Human Factors in Computer Systems*, 56 – 62.

Carder, H. P., Handley, S.J., & Perfect, T. J. (2004). Deconstructing the Tower of London: Alternative moves and conflict resolution as predictors of task performance. *The Quarterly Journal of Experimental Psychology, 57A*(8), 1459 – 1483.

Carlin, D., Bonerba, J., Phipps, M., Alexander, G., Shapiro, M., & Grafman, J. (2000). Planning Impairments in frontal lobe dementia and frontal lobe lesion patients. *Neuropsychologia, 38*, 655 – 665.

Carroll, J. M., & Rosson, M. B. (1987). The paradox of the active user. In Carroll, J.M. (Ed.), Interfacing thought: Cognitive aspects of human-computer interaction. Cambridge, MA: The MIT Press.

Cary, M., & Carlson, R. A. (2001). Distributing Working Memory Resources During Problem Solving. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27*(3), 836 – 848.

Cary, M., & Carlson, R. A. (1999). External support and the development of problem-solving routines. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25(4),* 1053 – 1070.

Charness, N. (1981). Search in Chess: Age and skill differences. *Journal of Experimental Psychology: Human Perception and Performance, 7(2),* 467 – 476.

Cheng, P. C-H. (1996). Scientific discovery with law-encoding diagrams. *Creativity Research Journal,* 9, 145-162.

Cheng, P. C.-H. (2002). Electrifying diagrams for learning: principles for effective representational systems. *Cognitive Science, 26(6),* 685-736.

Child, D. A. (1999). The effects of system response time on user behavior in a hypermedia environment. *Journal of Educational Multimedia and Hypermedia, 8(1),* 65 – 88.

Chronicle, E.P., MacGregor, J.N., & Ormerod, T.C. (2004). What makes an insight problem? The roles of heuristics, goal conception and solution recoding in knowledge-lean problems *Journal of Experimental Psychology: Learning, Memory & Cognition, 30,* 14-27.

Cohen, G. (1996). *Memory in the Real World.* Mahwah, NJ: Lawrence Erlbaum.

Corley, M. R. (1976). Pragmatic information processing aspects of graphically

accessed computer-aided design. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-6,* 434-439.

Crease, M.G., Brewster, S.A. (1998). Making progress with sounds – the design and evaluation of an audio progress bar. In Proceedings of ICAD'98 (Glasgow, UK), British Computer Society.

Davies, S. P. (2003). Initial and concurrent planning in solutions to well-structured problems. *The Quarterly Journal of Experimental Psychology, 56*(7), 1147 – 1164.

Davis, S., & Wiedenbeck, S. (1998). The effect of interaction style and training method on end user learning of software packages. *Interacting with Computers, 11*, 147 – 172.

DeGroot, A. D. (1965). *Thought and Choice in Chess*. The Hague, The Netherlands: Mouton.

Delaney, P. F., Ericsson, K. A., & Knowles, M . E. (2004). Immediate and sustained effects of planning in a problem solving task. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*(6), 1219 – 1234.

Emurian, H. (1991). Physiological responses during data retrieval: Consideration of constant and variable system response times. *Computers in Human Behaviour, 7*, 291 – 310.

Ericsson, K. A. (1974a). *Problem-solving behaviour with the 8-puzzle I: Time to solution.* Report No. 431. Department of Psychology, University of Stockholm.

Ericsson, K. A. (1974b). *Problem-solving behaviour with the 8-puzzle II: Distribution of latencies.* Report No. 432. Department of Psychology, University of Stockholm.

Ericsson, K. A. (1974c). *Problem-solving behaviour with the 8-puzzle III: Process in*

terms of latencies. Report No. 433. Department of Psychology, University of Stockholm.

Ericsson, K. A. (1975). *Instruction to verbalise as a means to study problem-solving processes with the 8-puzzle: A preliminary study.* Report No. 458. Department of Psychology, University of Stockholm.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task. *Annual Review of Psychology, 47*, 273-305.

Ernst, G. W., & Newell, A. (1969). GPS: A case study in generality and problem solving. Orlando, FL: Academic Press.

Geddes, B. W., & Stevenson, R. J. (1997). Explicit learning of a dynamic system with a non-salient pattern. *The Quarterly Journal of Experimental Psychology, 50A*(4), 742 – 765.

Gibson, J.J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), Perceiving, Acting and Knowing. Hillsdale, NJ: Erlbaum.

Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103(4),* 650-669.

Gilhooly, K. J., Phillips, L. H., Wynn, V., Logie, R. H., & Della Salla, S. (1999). Planning processes and age in the five-disc Tower of London task. *Thinking and Reasoning, 5*(4), 339 – 361.

Gilhooly, K. J., Wynn, V., Phillips, L. H., & Della Sala, S. (2002). Visuo-spatial and verbal working memory in the five-disc Tower of London task: An individual differences approach. *Thinking and Reasoning, 8*(3), 165 – 178.

Gobet, F., & Simon, H. A. (1996). Templates in Chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1 – 40.

Golightly, D. (1996). Harnessing the Interface for Domain Learning. *CHI Conference Companion*, 327 – 38.

Golightly, D., & Gilmore, D. (1997). Breaking the rules of direct manipulation. *Proceedings of the INTERACT.97 International Conference on Human.Computer Interaction,* 156.163. London: Chapman & Hall.

Gray, W. D., & Fu, W. (2001). Ignoring Perfect Knowledge in-the-world for imperfect knowledge in-the-head: Implications of Rational Analysis for Interface Design. *CHI Letters, 3*(1), 112-119.

Greeno, J. G. (1974). Hobbits and Orcs: Acquisition of a Sequential Concept. *Cognitive Psychology, 6*, 270 – 292.

Grossberg, M., Wiesen, R. A. & Yntema, D. B. (1976). An experiment on problem solving with delayed computer responses. *IEEE Transactions on Systems, Man, and Cybernetics, 3*, 219-222.

Gunzelmann, G., & Anderson, J. R. (2003). Problem solving: Increased planning with practice. *Cognitive Systems Research, 4,* 57 – 76.

Guynes, J. L. (1988). Impact of system response time on state anxiety. *Communications of the ACM, 31(3)*, 342 – 347.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition, 28*, 249 – 276.

Hayes, J. R., & Simon, H. A. (1974). Understanding written problem instructions. In L. W. Gregg (Ed.) *Knowledge and Cognition.* (pp. 167 – 200). Potomac, MD: Erlbaum.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3,* 275 – 310.

Hutchins, E., Hollan, J., & Norman, D. A. (1986). Direct manipulation interfaces. In D. A. Norman & S. W. Draper (Eds.) *User Centered System Design,* (pp. 87-124). Lawrence Erlbaum Associates Inc, Hillsdale, NJ.

Jeffries, R., Polson, P. G., Razran, L., & Atwood, M. E. (1977). A process model for missionaries-cannibals and other river-crossing problems. *Cognitive Psychology, 9,* 412 – 440.

Jones, G. (2003). Testing two cognitive theories of insight. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*(5), 1017 – 1027.

Karat, J. (1982). A model of problem solving with incomplete constraint knowledge. *Cognitive Psychology, 14,* 538 – 559.

Kershaw, T. C., & Ohlsson, S. (2001). Training for insight: The case of the nine-dot problem. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 489- 493). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W., & Greeno, J. G. (1985) Understanding and solving word arithmetic problems. *Psychological Review, 92(1),* 109 – 129.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 415 – 452.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1 – 48.

Knowles, M. E., & Delaney, P. F. (2005). Lasting reductions in illegal moves following an increase in their cost: Evidence from river crossing problems. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31,* 670 – 682.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17,* 248 – 294.

Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology, 22,* 143 – 183.

Krug, S. (2000). *Don't make me think: A common sense approach to web usability.* New Riders Publishing.

Laird, J., Newell, A., Rosenbloom, P. (1987). Soar: an architecture for general intelligence. *Artificial Intelligence,* 33, 1-64.

Lambert, G. N. (1984). A comparative study of system response time on program developer productivity. *IBM System Journal, 23,* 36 – 43.

Larkin, J. H. (1989). Display-based problem solving. In David Klahr, Kenneth Kotovsky, (Eds.) *Complex information processing: The impact of Herbert A. Simon.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ: p. 319-341.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65 – 99.

Lewis, A. B., & Mayer, R. E. (1987) Students' Miscomprehension of Relational Statements in Arithmetic Word Problems. *Journal of Educational Psychology, 79(4),* 363 – 371.

Long, J. (1976). Effects of delayed irregular feedback on unskilled and skilled keying performance, *Ergonomics, 19(2),* 183 – 202.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54,* (6, Whole No. 248).

Luchins, A. S., & Luchins, E. H. (1990). Task Complexity and Order Effects in Computer Presentation of Water-Jar Problems. *The Journal of General Psychology, 118(1),* 45 – 72.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information-processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27,* 176-201.

Maglio, P. P., & Kirsh, D. (1996). Epistemic Action Increases with Skill. *In Proceedings of the 18th Annual Conference of the Cognitive Science Society,* 391 – 396.

Martin, G. L., & Corl, K.G. (1986). System response time on user productivity. *Behaviour and Information Technology, 5(1),* 3 – 13.

Mayes, R.T., Draper, S.W., McGregor, A.M. and Oatley, K. (1988) Information flow in a user interface: The effect of experience of and context on the recall of MacWrite screens. In Jones, D.M. and Winder, R. (Eds.) *People and computers IV (pp. 257-289).* Cambridge: Cambridge University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Miller, R. B. (1968). Response time in man-computer conversational transactions. *Proceedings of the Joint Computer Conference,* 267 – 277.

Miller, G.A., Galanter, E., & Pribram, K.H. (1960). *Plans and the Structure of Behavior.* New York: Holt, Rinehart & Winston.

Morgan. P. (2005). "Now, where was I?" A cognitive experimental analysis of the influence of interruption on goal-directed behaviour. Unpublished D. Phil. thesis, Cardiff University.

Myers, B. A. (1985). The importance of percent-done progress indicators for computer-human interfaces. *Proc. ACM CHI'85 Conference.* (San Francisco, CA, 14-18 April), 11-17.

Neth, H., & Payne, S.J. (2002). Thinking by Doing? Epistemic Actions in the Tower of Hanoi. In W.D. Gray and C.D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 691-696). Mahwah, NJ: Lawrence Erlbaum.

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

Nielsen, J. (1994). *Usability Engineering.* Academic Press Inc.(London) Ltd.

Nilsson, N. (1971). *Problem-solving Methods in Artificial Intelligence.* New York, New York: McGraw-Hill.

Norman, D. A. (1988). *The Psychology of Everyday Things.* Basic Books Inc.

Noyes, J. M., & Garland, K. J. (2003). Solving the Tower of Hanoi: does mode of presentation matter? *Computers in Human Behaviour, 19,* 579 – 592.

O'Hara, K. P., & Payne, S. J. (1998). The Effects of Operator Implementation Cost on Planfulness of Problem Solving and Learning. *Cognitive Psychology, 35,* 34 – 70.

O'Hara, K., & Payne, S. J. (1999). Planning and the User Interface: The Effects of Lockout time and Error Recovery Cost. *International Journal of Human-Computer Studies, 50,* 41-59.

Ormerod, T.C. (2005). Planning and ill-defined problems. Chapter in R. Morris and G. Ward (Eds.) *The Cognitive Psychology of Planning*. London: Psychology Press.

Payne, S. J. (1991). Display-based action at the user interface. *International Journal of Man-Machine Studies, 35*, 275 – 289.

Payne, S. J., Howes, A., & Reader, W. R. (2001). Adaptively distributing cognition: a decision-making perspective on human-computer interaction. *Behaviour and Information Technology, 20,* 339-346.

Phillips, L. H., Gilhooly, K. J., Logie, R. H., Della Salla, S., & Wynn, V. E. (2003) Age, working memory, and the Tower of London task. *European Journal of Cognitive Psychology, 15*(2), 291 – 312.

Phillips, L. H., Wynn, V., Gilhooly, K. J., Della Salla, S., & Logie, R. H. (1999). The role of memory in the Tower of London task. *Memory, 7*(2), 209 – 231.

Pizlo, Z., & Li, Z. (2005). Solving Combinatorial Problems: The 15-puzzle. *Memory & Cognition, 33(6)*, 1069 – 1084.

Ramsay, J. Barbesi, A. and Preece, J. (1998). A psychological investigation of long retrieval times on the World Wide Web. *Interacting with Computers, 10,* 77-86.

Ratterman, M. J., Spector, L., Grafman, J., Levin, H., & Harward, H. (2001). Partial and total-order planning: evidence from normal and pre-frontally damaged populations. *Cognitive Science, 25*, 941 – 975.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: evidence from eye movements. *Psychological Science, 12*(1), 48 – 55.

Robertson, S. P., & Black, J. B. (1986). Structure and development of plans in computer text editing. *Human-Computer Interaction, 2*, 201–226.

Rose, G. M., Lees, J. & Meuter, M. (2001). A refined view of download time impacts upon e-consumer attitudes and patronage intentions toward e-retailers. *The International Journal of Media Management, 3(2)*, 105 – 111.

Sacerdoti, E., (1975). The Non-linear Nature of Plans. *International Joint Conference on Artificial Intelligence*, 206 – 214.

Schar, S. G. (1996). The Influence of the user interface on solving well- and ill-defined problems. *International Journal of Human-Computer Studies, 44*, 1 – 18.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London (Series B), B298*, 199-209.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7*, 268-288.

Simon, H. A., & Lea, G. (1974) Problem solving and rule induction: A unified view. In L.W. Gregg (Ed.), *Knowledge and cognition, (pp. 105-127)*. Hillsdale, NJ: Erlbaum.

Stokes, M. T., Halcomb, C. G., & Slovacek, C. P. (1988). Delaying user responses to computer-mediated items enhances test performance. *Journal of Computer-Based Instruction, 15*, 99-103.

Svendsen, G. B. (1991). The influence of interface style on problem solving. *International Journal of Man-Machine Studies, 35*, 379 – 397.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General, 119*(2), 176 – 192.

Sweller, J., Mawer, R. B., & Ward, M. R. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 112*(4), 639 – 661.

Teal, S. L. & Rudnicky, A. I. (1992). A performance model of system delay and user strategy selection. *Human Factors in Computer Systems, ACM*, 295 – 305.

Thevenot, C., & Oakhill, J. (2005). The strategic use of alternative representations in arithmetic word problem solving. *The Quarterly Journal of Experimental Psychology, 58A,* 1311 – 1323.

Thomas, J. C. (1974). An analysis of behavior in the Hobbits-Orcs problem. *Cognitive Psychology, 6*, 257 – 269.

Trimmel, M., Meixner-Pendleton, M., & Haring, S. (2003). Stress response caused by system response time when searching for information on the Internet. *Human Factors, 45(4),* 615-621.

Trudel, C., & Payne, S. (1995). Reflection and Goal management in Exploratory Learning. *International Journal of Human-Computer Studies, 42*, 307 – 339.

Unterrainer, J. M., Rahm, B., Kaller, C. P., Leonhart, R., Quiske, K., Hoppe-Seyler, K., Meier, C., Muller, C., & Halsband, U. (2004). Planning abilities and the tower of London: is this task measuring a discrete cognitive function? *Journal of Clinical and Experimental Neuropsychology, 26*(6), 846 – 856.

VanLehn, K. (1996). Cognitive Skill Acquisition. *Annual Review of Psychology, 47*, 513 – 539.

VanLehn, K. (1991). Rule acquisition events in the discovery of problem-solving strategies. *Cognitive Science, 15*, 1 – 47.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.) *Foundations of Cognitive Science* (pp. 526-579). Cambridge, MA: M.I.T. Press.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75 – 100.

Ward, G. (1993). An experimental investigation of executive processes. Unpublished D. Phil. thesis, University of Oxford.

Ward, G., & Allport, A. (1997) Planning and Problem-solving using the Five-disc Tower of London Task. *The Quarterly Journal of Experimental Psychology, 50A(1),* 49 – 78.

Weisberg, R.W., & Alba, J.W. (1981a). An examination of the alleged role of 'fixation' in the solution of several 'insight' problems. *Journal of Experimental Psychology: General, 110(2),* 169-192.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21(2)*, 179-217.

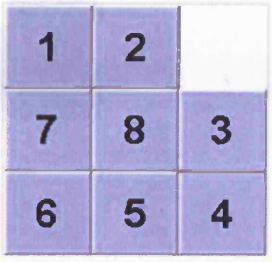
Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*, 87 – 122.

Zhang, J., & Wang, H. (2005). The effect of external representations on numeric tasks. *Quarterly Journal of Experimental Psychology, 58A(5)*, 817-838.

Zona Research. (2001). The Need for Speed 2. *Zona Market Bulletin, 5*, 1 – 9.

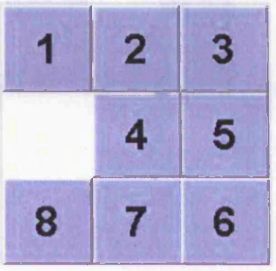
Appendix A: Practice Trials, Short Solution Path Start States and Long Solution Path Start States for Experiment 3

Practice State 1



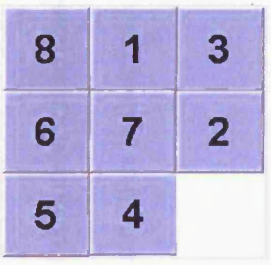
6 Moves

Practice State 2



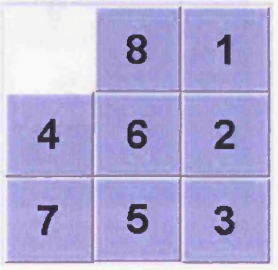
5 Moves

SH1 Problem Start State



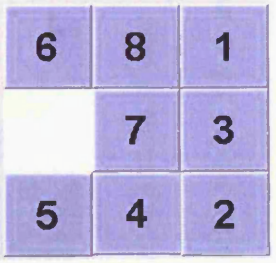
12 Moves

SH2 Problem Start State



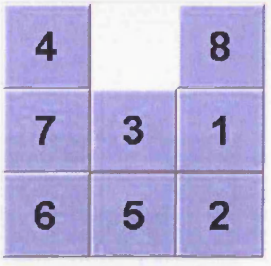
12 Moves

LG1 Problem Start State



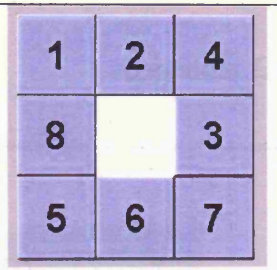
17 Moves

LG2 Problem Start State



17 Moves

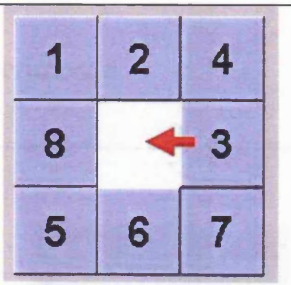
Appendix B: Copy of Hint given to all 8-puzzle Subjects from Experiments 3 & 4



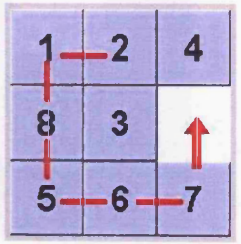
- Consider the above arrangement of tiles
- Tiles 1 & 2 are in place but tiles 3 & 4 are reversed
- In order to get tile 3 in place it will be necessary to moves tiles 1 & 2

In general it may be better to focus on getting tiles in their correct numerical sequences rather than always placing tiles in their individual places.

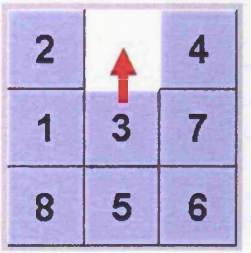
So…



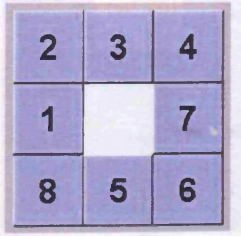
- Move 3 Over



- Rotate the sequence '765812' anti-clockwise

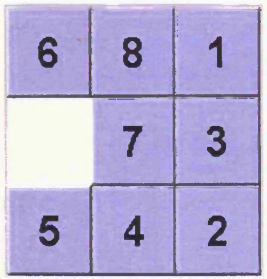


- Move 3 into position

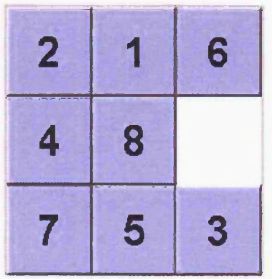


- Tiles 1, 2, 3 & 4 are now in the correct numerical order.
- The goal state is now much closer

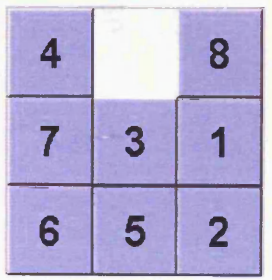
Appendix C: The 8 Different 17-move Problem Start States with Solution Paths used in Experiment 4



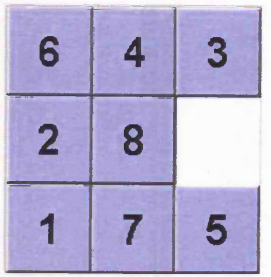
6,8,1,3,2,4,5,6,7,2,4,5,6,7,8,1,2



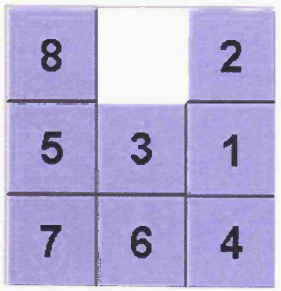
6,1,8,6,3,5,6,4,2,8,1,3,4,2,8,1,2



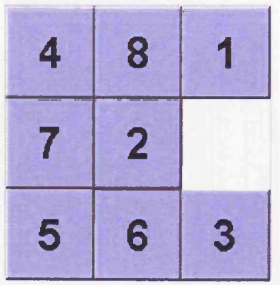
8,1,2,5,6,7,4,8,1,2,3,4,8,1,2,3,4



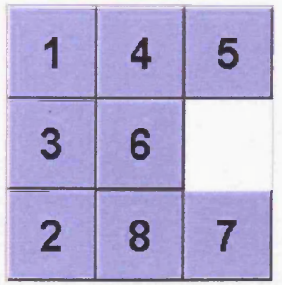
5,7,8,4,6,2,1,8,7,5,4,6,2,1,8,7,6



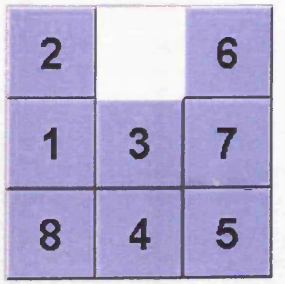
2,1,3,5,7,6,5,2,1,3,4,5,6,7,8,1,2



2,6,5,7,4,8,1,2,3,5,6,4,8,1,2,3,4



6,3,2,8,7,6,5,4,3,2,8,7,6,5,4,3,2



3,7,5,4,7,5,6,3,2,1,8,7,5,6,4,5,6

Appendix D: Mental Arithmetic Screener Used in Experiment 5

Q1.

Calculate the sum of:

(18 x 9) / 2

Q2.

Calculate the sum of:

(8 x 23) x 2

Q3.

Calculate the sum of:

(2886 / 6) + 30

Q4.

If the Loch Ness monster's length is twenty yards shorter than twice her length, how long is she?

Appendix E: Experiment 5 Water Jars Specifications and Solution Paths

LEVEL	Solution Length	Jar Sizes	Goal State	Start State	Solution Path
1	3 Moves	6 / 4 / 3	4 / 0 / 2	6 / 0 / 0	(A – B) (A – C) (B – A)
2	3 Moves	5 / 4 / 3	4 / 1 / 0	5 / 0 / 0	(A – B) (B – C) (C – A)
3	4 Moves	9 / 7 / 2	5 / 4 / 0	9 / 0 / 0	(A – C) (C – B) (A – C) (C – B)
4	4 Moves	8 / 5 / 3	2 / 5 / 1	8 / 0 / 0	(A – C) (C – B) (A – C) (C – B)
5	5 Moves	11 / 6 / 5	4 / 6 / 1	11 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (A – B)
6	5 Moves	10 / 5 / 4	4 / 5 / 1	10 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (A – B)
7	6 Moves	10 / 7 / 3	9 / 0 / 1	10 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (C – A) (B – C)
8	6 Moves	11 / 8 / 3	2 / 8 / 1	11 / 0 / 0	(A – C) (C – B) (A – C) (C – B) (A – C) (C – B)

Appendix F: Experiment 6 Water Jars Specifications and Solution Paths

LEVEL	Solution Length	Jar Sizes	Goal State	Start State	Solution Path
1 (Practice)	3 Moves	6 / 4 / 3	4 / 0 / 2	6 / 0 / 0	(A – B) (A – C) (B – A)
2	4 Moves	8 / 5 / 3	2 / 5 / 1	8 / 0 / 0	(A – C) (C – B) (A – C) (C – B)
3	4 Moves	15 / 7 / 3	11 / 1 / 3	15 / 0 / 0	(A – B) (B – C) (C – A) (B – C)
4	5 Moves	11 / 6 / 5	7 / 0 / 4	11 / 0 / 0	(A – C) (C – B) (A – C) (C – B) (B – A)
5	5 Moves	11 / 6 / 5	4 / 6 / 1	11 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (A – B)
6	6 Moves	10 / 7 / 3	9 / 0 / 1	10 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (C – A) (B – C)

Appendix G: Experiment 7 Water Jars Specifications and Solution Paths

LEVEL	Solution Length	Jar Sizes	Goal State	Start State	Solution Path
1 (Practice)	3 Moves	6 / 4 / 3	4 / 0 / 2	6 / 0 / 0	(A – B) (A – C) (B – A)
2 (Practice)	3 Moves	5 / 4 / 3	4 / 1 / 0	5 / 0 / 0	(A – C) (C – B) (A – C) (C – B)
3	4 Moves	8 / 5 / 3	2 / 5 / 1	8 / 0 / 0	(A – C) (C – B) (A – C) (C – B)
4	4 Moves	15 / 7 / 3	11 / 1 / 3	15 / 0 / 0	(A – B) (B – C) (C – A) (B – C)
5	6 Moves	10 / 7 / 3	9 / 0 / 1	10 / 0 / 0	(A – B) (B – C) (C – A) (B – C) (C – A) (B – C)
6	7 Moves	12 / 6 / 5	3 / 4 / 5	12 / 0 / 0	(A – C) (C – B) (A – C) (C – B) (B – A) (C – B) (A – C)

Appendix H: Screener used in Experiments 6 and 7

Q1.

Calculate the sum of:

(18 x 9) / 2

Q2.

Calculate the sum of:

(17 x 8) x 2

Q3.

Calculate the sum of:

(2886 / 6) + 11

Q4.

If the Loch Ness monster's length is twenty yards shorter than twice her length, how long is she?

Hint: It may be useful to think of this question in terms of an algebraic formula