

Deconstructing the Tower of London: Alternative moves and conflict resolution as predictors of task performance

Hassina P. Carder, Simon J. Handley, and Timothy J. Perfect

University of Plymouth, Plymouth, UK

Despite widespread use the cognitive demands of the five-disc Tower of London (TOL) are unknown. Research suggests that conflict moves (those that are essential to the solution but do not place a disc in its final position) are a key aspect of performance. These were examined in three studies via a verification paradigm, in which normal participants were asked to decide whether a demonstrated move was correct. Experiment 1 showed that individual move latencies increase with the number of intermediate moves until the disc is placed in its goal position (resolution). Post hoc tests suggested that the number of alternative moves and moves to resolve a disc were independent predictors of performance. Experiment 2 successfully manipulated these factors in an experimental design. Experiment 3 showed that they remain determinants of performance as familiarity increased. Overall, errors on the task were significantly correlated with spatial memory. The implications of these findings for the use of the TOL in cognitive psychology and as an assessment tool are discussed.

The Tower of London (TOL; Shallice, 1982), a variation of the Tower of Hanoi (TOH), is a task that has been the focus of research within both cognitive psychology and clinical neuropsychology. The task uses three pegs (sometimes of different lengths) and a number of equal-sized coloured discs. The participant is shown a start and goal configuration of discs and has to transform the problem by moving one disc at a time. Each problem should be solved in the minimum number of moves. The three-disc TOL is suitable for use with various clinical populations but is too simple for normal participants, so a five-disc version has been developed (Ward & Allport, 1997) and is used in the current series of studies. It is illustrated in Figure 1.

Despite widespread use, little is known about the precise cognitive processes involved in TOL performance. One common claim is that TOL performance reflects the efficiency of

Correspondence should be addressed to Hassina Carder, Department of Psychology, University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA, UK. Email: hcarder@plymouth.ac.uk

To Simon Paice and Lynne James for writing the computer programs. Thanks also to Geoff Ward and Louise Phillips for comments on an earlier draft. This research is funded by an ESRC grant to the first author No: R42200034072.

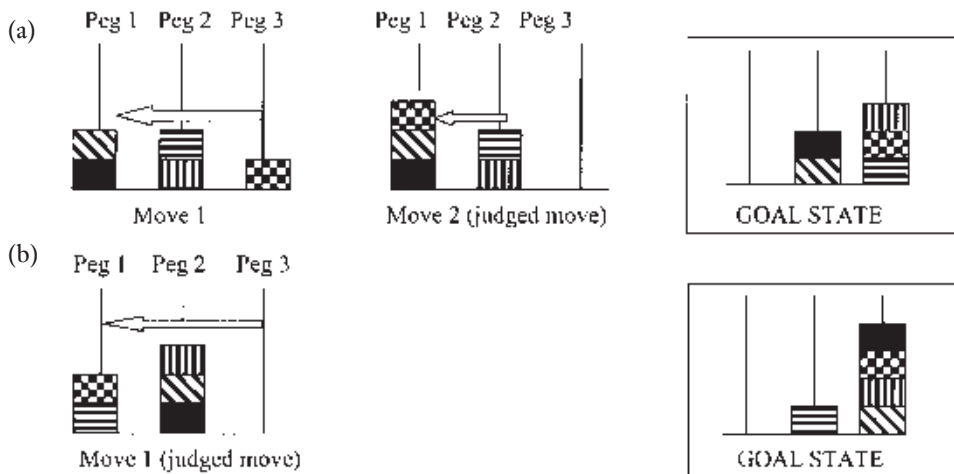


Figure 1. Illustration of the resolution factor in the verification TOL. The arrow illustrates the movement of disc that the participant is asked to judge. In both examples a confirm move is illustrated, and the decision is required after the move has been completed. This move may occur in the middle of a solution. Therefore where appropriate the previous moves are shown. Figure 1a is a one-resolution problem because the chequered disc is placed in its goal position (resolved) after one intermediate move. This intermediate move is the movement of the horizontally striped disc from Peg 2 to Peg 3. In Figure 1b the chequered disc is resolved after three intermediate moves. These are: (1) vertical striped disc moves from Peg 2 to Peg 1; (2) diagonal striped disc moves from Peg 2 to Peg 3; (3) the vertical striped disc moves from Peg 1 to Peg 3. Then the chequered disc can be placed in its goal position on Peg 3.

planning processes. Typically participants are instructed to plan the whole sequence of moves before executing their solution. Clinical populations such as frontal lobe patients have shown impaired performance compared to normal controls, and this has been interpreted as an inability to plan efficiently (e.g., Owen, Downes, Sahakian, Polkey, & Robbins, 1990; Shallice, 1982, 1988). However, this interpretation has proved problematic. For example, manipulations that reduce the opportunity for planning, such as limiting the time for planning a solution, do not impair performance (Phillips, Wynn, Gilhooly, Della Sala, & Logie, 1999). There are also problems in providing a clear and unambiguous definition of what planning involves. These issues have led to the suggestion that characterising TOL as a planning task is unproductive (Miyake et al., 2000b; Phillips, Wynn, McPherson, & Gilhooly, 2001).

An alternative argument is that problem difficulty is a function, not of planning efficiency, but of the ability to successfully inhibit inappropriate move selections at specific points within a solution path. These demands are at their greatest during “conflict moves”, where there is an apparent mismatch between the end goal of the problem and a current subgoal. During a conflict move the disc is placed in an intermediate position during the solution path. This contrasts with a forward move where a disc is placed directly into its final position. Conflict moves are essential to solve many problems in the minimum number of moves, and participants appear to experience difficulty with them (e.g., Goel & Grafman, 1995; Morris, Miotto, Feigenbaum, Bullock, & Polkey, 1997). Many researchers have argued that these moves, which superficially take you away from the goal state but in fact lead to the most efficient solution, are difficult because they initially appear counterintuitive (e.g., Goel & Grafman, 1995; Ward & Allport, 1997). Goel and Grafman argue that in TOH the problem that frontal

patients have with conflict moves is not due to a general decline in memory or difficulties in planning, but is because conflict moves require the inhibition of a prepotent (but inappropriate) move in favour of an alternative but appropriate response.

Goel and Grafman (1995) argue that frontal patients cannot see behind the “trick” of a conflict move and so fail to see that moves other than the most obvious or prepotent (in stimulus-driven terms) will result in a more efficient solution. They argue that this performance deficit is independent of a planning interpretation and is consistent with a class of explanation that explains frontal lobe deficits on various inhibition tasks including the antisaccade and the Stroop test. Likewise Miyake et al. (2000b) noted that the performance of normal participants often fails at the first conflict move on TOH, and this results in longer solution paths. Ward and Allport (1997) have shown in normal participants that the latency and errors for preparation time (i.e., time taken to plan a move sequence) on TOL is best predicted by the number of chunks of consecutive conflict moves (i.e., those that transfer discs from and back to the same peg). One of the major aims of this paper is to examine the specific characteristics of conflict moves that contribute to their difficulty.

The claim that TOL performance reflects the efficiency of inhibitory processes is consistent with the view that the task measures executive functioning, in which inhibitory control is seen as a core function (Miyake et al., 2000b). However, there are a number of problems in using a task that is poorly understood to examine a construct that itself is poorly defined. Descriptions of executive control processes in cognition have created much interest and debate in recent years (e.g., Baddeley, 1998; Miyake, Emerson, & Friedman, 2000a; Parkin, 1998) but there are several issues that have impaired the clear specification of the functions of the executive system, some of which relate to executive tasks themselves. First is a lack of consensus about which tasks or measures to use in theoretical and clinical settings. Second, many executive tasks lack construct validity, and the processes involved in performance are poorly understood. A third issue is that executive tasks are complex, and impairment can be a function of a number of factors (Miyake et al., 2000a).

The problem of defining executive functions via performance on executive tasks is clearly illustrated by the TOL. As we have seen there is disagreement over whether the task is best thought of as a measure of planning or inhibition. In addition, patients with injury to the prefrontal cortex, a brain area commonly associated with executive functioning, do not always show impaired performance on the task (see, e.g., Shallice, 1982, 1988). This could be due to the complexity of the TOL and/or it could be due to the heterogeneity of frontal deficits, which tend to produce a more diverse range of behavioural impairment than that caused by damage in other areas of the brain (Parker & Crawford, 1992). The TOL has also been reported to have low reliability, and this could complicate the interpretation of findings (Schnirman, Welsh, & Retzlaff, 1998). This might in part be due to the fact that executive involvement is believed to be strongest when a task is novel, and therefore repeated encounters can lead to practice effects, which may reduce reliability.

It is important that the cognitive processes involved in the TOL are fully understood so that deficits in performance can be associated with deficits in specific cognitive processes. Hence this paper addresses these issues by examining the factors that affect TOL difficulty in a set of items often used in the literature (e.g., Gilhooly, Phillips, Wynn, Logie, & Della Sala, 1999; Phillips et al., 1999). We show that the specific characteristics of the items used are of fundamental importance in determining performance, and that the difficulty of particular

moves cannot be explained with recourse to simple one-word descriptions, such as planning or inhibition.

The TOL is typically administered by providing participants with a start and a goal state and asking them to solve the problem by moving discs. The most frequently used dependent variables are the number of problems solved in the minimum number of moves within 60 seconds (Shallice, 1982), the number of excess moves over a fixed number of trials, or global latency measures (e.g., Owen et al., 1990). Participants generate their own solution paths, and these may vary between individuals particularly given that some problems have a number of equally efficient solutions. The TOL is a complex task, and the use of global measures of performance arguably only provides a very gross reflection of the efficiency of executive processes. We have addressed this issue in the present series of experiments by adopting a verification paradigm in which participants have to make speeded judgements about the appropriateness of a particular move within a TOL solution.

The verification paradigm has not been employed in TOL administration before now, but has several advantages over the standard paradigm of TOL administration. For instance, it allows for measurement of individual moves, and this enables fine-grain analysis of performance. Indeed Ward and Allport (1997) argued that analysis at the level of the individual move can provide a rich source of evidence by which to evaluate performance. Likewise Morris et al. (1997) manipulated the characteristics of the first move and argued that this clarified the nature of the impairments between different frontal groups. However, in our verification paradigm it is possible to experimentally manipulate task characteristics at positions other than the first move, and this second advantage guarantees that the experimenter has control over the solution path and standardizes the moves that the participant experiences. A third advantage of the verification paradigm is that the task complexity is broken down. In executive tasks a number of processes may be required for satisfactory performance over an extended period of time (Shallice & Burgess, 1991). Whilst it may be valid to assess executive functioning with tasks that share this everyday characteristic, it may also be informative to use more sensitive measures in existing clinical tests, and the verification paradigm enables this by reducing overall performance on TOL into chunks.

As we have seen, Goel and Grafman (1995) have argued that conflict moves on TOH are difficult because conflict moves require the inhibition of a prepotent but inappropriate response in favour of an appropriate alternative. Welsh, Satterlee-Cartmell, and Stine (1999) have also argued that inhibition ability is related to TOL performance. In our view this characterization is oversimplistic. Although conflict moves, by their very definition, require the satisfaction of a conflict between different possible moves, they also have a number of other characteristics that may contribute to their difficulty. For example, a feature of a forward move is that it is resolved immediately—that is, the disc is immediately placed in its goal position. In contrast in a conflict move there are a number of intermediate moves that are necessary before the disc is finally placed. In order to be certain that a move is appropriate it is therefore necessary to think ahead and to consider at what point the focal disc will be placed in the goal position, following the intermediate move(s). This analysis suggests that conflict moves are not a homogeneous category, but will vary in difficulty as a function of the number of intermediate moves required. In this paper we refer to this characteristic as the resolution gap. In Experiment 1 we used the verification paradigm to manipulate this resolution gap to see how it affects performance. Moves are defined as a forward move if the resolution gap is zero. In a

one-resolution trial the conflict move is taken, there is one intermediate move, and then the focal disc is placed in its final position, as illustrated in Figure 1a. A three-resolution problem has three intermediate moves and is illustrated in Figure 1b, and a five-resolution problem has five intermediate moves.

Research suggests that TOL/TOH are executive tasks that engage both working memory (WM) and inhibition processes (e.g., Goel & Grafman, 1995; Miyake et al., 2000b; Owen et al., 1990; Welsh et al., 1999; Owen, Doyon, Petrides, & Evans, 1996). There is plenty of evidence that spatial working memory (SWM) is involved in performance in TOH (e.g., Handley, Capon, Copp, & Harper, 2002) and TOL (Gilhooly, Wynn, Phillips, Logie, & Della Sala, 2002; Phillips et al., 1999). Given the likely role of WM in the task, in addition to the experimental manipulations, Experiment 1 employs an individual difference measure of SWM as measured by a passive TOL span task (adapted from Morris et al., 1997) alongside the experimental design.

In summary, Experiment 1 uses a previously unused verification paradigm to examine performance in TOL at the level of the individual move. It explores the nature of conflict moves by manipulating the resolution gap of individual moves on TOL, so providing a comparison of performance between forward and conflict moves of predicted increasing difficulty.

EXPERIMENT 1

Method

Participants

A total of 47 undergraduates of the University of Plymouth participated for course credit. All participants had normal colour vision but none were familiar with TOL. The data from one participant was dropped as they were not administered all 32 problems.

Procedure

Participants completed the tasks in one testing session of approximately 30 minutes. They were given the verification task first and then the span task. What follows is a summary of each task used in Experiment 1.

Tower of London task. A computerized version of the five-disc, three-peg TOL was used (as developed for normal participants by Ward & Allport, 1997). One practice trial followed by 32 problems was administered in a unique random order to each participant. The participant was told that the goal state was situated in the upper portion of the screen and that the lower portion contained the start state. They were required to judge whether particular moves were correct or incorrect in order to solve the problem in the minimum number of moves (which they were not provided with). They were told that there were two rules to the task: (1) Only the top disc on any peg can be moved, and (2) only one disc at a time can be moved. The computer performed the moves up to and including a crucial move. No preplanning time was allowed. Each move took 2,500 ms to complete. Immediately after the crucial move had been demonstrated the computer asked the participant to judge whether it was correct or not by clicking the appropriate on-screen button reading "correct" or "incorrect". In order that the participant was ready to respond, the computer indicated via an on-screen message the number of remaining moves until the

judgement was required. Participants were told that their responses were timed and that they should work as quickly as possible, while avoiding errors.

Half the problems demonstrated the correct move and half an incorrect move. The confirm judgements consisted of four problems where the move to be judged was resolved immediately (i.e., forward moves). There were four problems in which the disc was resolved on the second move (one intermediate move), four on the fourth move (three intermediate moves), and four on the sixth move (five intermediate moves), as described above. There are two possible pegs to which a disc can move, and so the dispute trials were matched to confirm trials by moving the disc onto the wrong peg. In addition the peg order and disc colours were changed, resulting in a problem that appeared different superficially but was logically identical. The computer recorded the number of errors and latency of participants' judgements in each of the 32 trials. Response latency (in milliseconds) for each judgement was measured from the time that the crucial move was finally placed until the participant had responded by clicking the on-screen button with a mouse.

Tower of London memory span. This task employed a modified version of the verification computer program featuring the five-disc, three-peg TOL (Ward & Allport, 1997). The task is based on a measure used by Morris et al. (1997), and is designed as a passive span measure of storage. No practice trials were issued. The participant witnessed the start state, but not the goal state, and was asked to watch the computer perform the necessary moves to solve the problem, starting with a two-move problem.

The computer took 2,500 ms to move a disc from its original to the destination peg, and meanwhile "the computer is making its moves" appeared on screen. When all moves had been demonstrated the apparatus returned to the start state, and "copy that sequence" appeared on screen. The participant then reproduced the series of moves using a drag and drop function with the mouse. When the participant had finished they were asked to press a "submit" button, which remained on screen throughout their turn. The participant was told via an on-screen message whether they had correctly completed each trial. Each trial consisted of a problem with between 2 and 10 moves. There were up to three attempts at each level, starting with a two-move problem and increasing by one additional move at each level. The participant had to reproduce at least one of these trials correctly before the difficulty increased by one additional move. Participants were warned via an on-screen message if the number of moves would increase. However, if the participant did not get any trial correct at a given level the task ended. A global score was calculated where one point was awarded for each trial successfully reproduced. The maximum score obtainable was 27.

Apparatus

All the tasks were presented on a Hewlett Packard Vectra PC with a 17" screen. The resolution was set at 800 x 600 high colour (16 bit), and tasks were presented via Windows 98. The participants interacted with the computer using a computer mouse.

Design

The resolution factor was in four levels (0, 1, 3, 5), and the decision type had two levels (confirm and dispute). There were two dependent variables: errors and latency in milliseconds for correct trials. A TOL span score was also obtained for correlation analysis with the factorial data.

Results

Experimental analysis

The reaction time data showed some positive skewing in the distribution, and so individual cells that contained values greater than twice the standard deviation were trimmed and were replaced with the condition mean for all cells (Ratcliff, 1993). In this and the subsequent experiments trimming did not alter the pattern of results observed.

A 4 (resolution) × 2 (confirm/dispute) analysis of variance (ANOVA) on the latency data showed a main effect of resolution, $F(3, 138) = 36.082, MSE = 1,127,330, p < .001, \eta^2 = .40$, indicating that increasing the number of moves until a disc is resolved increases latency. Pairwise comparisons using the LSD adjustment revealed a significant difference between zero-resolution and one-resolution problems ($p < .001$), while the increase of resolution between one and three approached significance ($p = .079$) as did the difference between three-resolution and five-resolution problems ($p = .081$). As Table 1 shows, the results reflect the fact that decision latency increased with the increase in resolution gap between zero and three moves, but the difference between three- and five-resolution problems was due to a reduction in latency. There was a main effect of decision type, $F(1, 46) = 7.870, MSE = 1,623,061, p = .007, \eta^2 = .146$, with dispute trials being significantly quicker than confirm trials. This may be a result of the fact that once a reason to dispute a move is found, continued processing is unnecessary, leading to a reduction in latency. There was no interaction between the two factors, $F(3, 138) = 1.192, MSE = 764,376, p = .315, \eta^2 = .025$.

For the error data a 4 × 2 within-subjects ANOVA showed a main effect of decision type, $F(1, 46) = 10.056, MSE = 2.645, p = .003, \eta^2 = .179$, revealing more accurate responses on dispute trials, suggesting a preference toward disputing a move during novel encounter. There was also evidence for an effect of resolution, $F(3, 138) = 18.489, MSE = 0.743, p < .001, \eta^2 = .287$ such that increasing the number of moves until a disc is resolved increased errors. Pairwise comparisons using a LSD method revealed that there were significant differences between zero-resolution and one-resolution gaps ($p < .001$), while an increase in the resolution gap between one-resolution and three-resolution approached significance ($p = .08$), indicating some increase in errors as the resolution gap increased between zero and three. However,

TABLE 1
The effects of resolution gap response latency^a for correct responses and number of errors for confirm and dispute trials in Experiment 1^b

Resolution	Confirm				Dispute			
	Response latency		Errors		Response latency		Errors	
	M	SD	M	SD	M	SD	M	SD
0	1,666	986	0.28	0.50	1,530	657	0.94	1.11
1	2,863	1426	1.66	1.36	2,589	1434	0.74	0.97
3	3,403	1674	2.09	1.36	2,873	1512	0.77	0.98
5	2,946	1542	1.69	1.34	2,411	1470	1.13	1.12

Note: Resolution is the number of intermediate moves before the disc is placed in its goal position.

^a In ms. ^b $n = 47$.

between the three-resolution and five-resolution conditions the difference was not significant ($p = .87$). There was a significant interaction between the resolution and decision factors, $F(3, 138) = 20.849$, $MSE = 0.822$, $p < .001$, $\text{Eta}^2 = .312$, such that on confirm trials the errors generally increased with the widening of the resolution gap between zero and three resolution, while on dispute trials there was a slight reduction in errors between zero and three resolution. Post hoc tests using Scheffe's test revealed all confirm judgements were significantly different from one another ($p < .05$) except between the five- and one-resolution judgement ($p > .05$). In dispute trials none of the conditions was significantly different from any other ($p > .05$). One way of interpreting this finding is that as an item becomes more difficult as a result of resolution gap, there is a tendency to dispute a move in some participants. This strategy will result in greater errors on confirm as a function of resolution gap. On trials in which a dispute response is correct, these participants will exhibit more accurate performance as problem difficulty increases. This will result in an accentuation of the effect of resolution for confirm trails, but an attenuation for dispute trials. This is exactly the pattern observed in the error data.

The findings of Experiment 1 provide partial confirmation that the difficulty of conflict problems is a function of the number of moves that lie between the onset move and the final placing of the disc. This suggests that participants are engaging in a form of on-line planning in which the assessment of a move is evaluated by looking ahead and considering subsequent moves that would allow the goal position for that disc to be achieved. Up to a point the data show that the more moves that need to be considered, the more difficult the problem. However, some effects were not as expected, most notably the drop in both latency and error rates between the three-resolution and five-resolution conditions. Although we adopted a verification paradigm with the aim of reducing task complexity there are nevertheless a range of problem features that may have influenced accuracy. For example, it is generally agreed that problems requiring a large number of moves are harder than shorter problems. Likewise given that participants are trying to find the most efficient solution it is likely that moves with a large number of legal alternative moves will be more difficult than moves in which there is only one other legal alternative move. Indeed the number of alternative move possibilities may reflect another key characteristic of conflict moves: that of competing move choice. In order to define the key characteristics of problem difficulty all judged moves were defined in terms of the total number of remaining moves in the problem from the crucial move until the end, the number of moves until the crucial disc is resolved (resolution), and the number of legal alternative moves. These are given in Table 2 and clearly show that the three-resolution problems had a greater number of alternative moves than the five-resolution problems, which could account for the reduced difficulty of the five-resolution problems compared to the three-resolution problems.

The characteristics of the judged moves were analysed using a multiple regression technique where the weakest predictor was removed, in order to see which factors best predicted performance. A backward elimination multiple simultaneous regression technique showed that 69.5% of the variance in latency data was predicted by the decision type, resolution gap, and alternative moves. As Table 3 shows these factors emerged as independent predictors of performance. For the error data the decision type and resolution gap proved to be the only independent predictors of performance, explaining 41% of the variance in errors. The alternatives factors was not correlated with the other predictors ($r = -.025$, $r = -.193$, $p > .05$). The resolution factor and the moves until the end of the problem were correlated with each other ($r = .858$, $p < .05$), which may suggest some instability in the results of the regression.

TABLE 2
 Characteristics of the crucial moves used in problems in Experiment 1

<i>Problem no. (confirm and dispute versions)</i>	<i>No. intermediate moves until disc is in goal position (resolution)</i>	<i>Moves until end of problem</i>	<i>No. legal alternative moves to that demonstrated</i>
1 and 17	0	3	1
2 and 18	0	3	1
3 and 19	0	3	1
4 and 20	0	3	1
5 and 21	1	5	1
6 and 22	1	4	5
7 and 23	1	4	1
8 and 24	1	6	1
9 and 25	3	5	5
10 and 26	3	8	3
11 and 27	3	5	3
12 and 28	3	6	5
13 and 29	5	7	1
14 and 30	5	9	1
15 and 31	6	7	1
16 and 32	5	8	3

However, importantly the alternatives variable was not correlated with either of the remaining predictors, and consequently the analysis gives an unambiguous indication for the importance of this variable in independently predicting performance.

Correlational analysis

We examined the relationship between the WM spatial span task and performance on the TOL verification paradigm. No consistent relationships were established with the latency data, but there were significant relationships with the error data. These revealed a relationship of $r = -.491$ ($p < .001$) between SWM errors on the task. This illustrates that better memory spans are associated with lower error rates on the verification task, and it confirms the role of SWM in the TOL.

TABLE 3
 Multiple regression for Experiment 1

	<i>Decision</i>		<i>Resolution</i>		<i>Alternatives</i>		<i>Model F value</i>	<i>R²</i>
	<i>Beta weight</i>	<i>t value</i>	<i>Beta weight</i>	<i>t value</i>	<i>Beta weight</i>	<i>t value</i>		
Latency	-.297	-2.83**	.587	5.51**	.429	4.03**	21.24, $p < .001$.695
Errors	-.428	-3.00**	.502	3.51**	.224	1.58	10.10, $p < .001$.411

Note: The nature of the decision (confirm/dispute), the resolution gap, the number of legal alternative moves, the number of moves in the problem, and the number of moves from the demonstrated move until the end of the problem were entered into the regression, and a backward elimination technique was used.

** significant at $p < .01$.

Discussion

Experiment 1 suggests that one reason that conflict moves are difficult is that it is necessary to think ahead in order to evaluate whether a move will solve the problem in the minimum number of moves. Hence forward problems are easiest because they are resolved immediately, and three-resolution problems are harder because they require the participant to track three additional moves, which is time consuming and likely to result in an error. We have demonstrated a significant negative correlation between errors and TOL span, suggesting that the task involves some form on-line manipulation of discs in WM. It is also worth noting that there was no significant correlation between memory span and latency data, suggesting that there is no relationship between speed at which participants can judge moves and their WM capacity. This finding is consistent with recent dual-task work, which shows that spatial secondary tasks increase the number of moves made, but do not impact on move latency (Phillips et al., 1999).

The evidence suggests that conflict moves are not a homogeneous category, but vary in difficulty as a function of resolution gap and, as the multiple regression analysis shows, the number of legal alternative moves. This suggests that at any given point in the solution path participants are considering a range of alternative moves, and that the greater the number of alternatives the longer the judgement takes. One interpretation of these findings is that the resolution, or “think ahead” component of conflict moves, maps on to some sort of on-line planning function, whereas deciding between alternative moves requires some form of inhibitory function, where inefficient moves must be resisted and appropriate choices acted upon. If this analysis is correct then each factor should make an independent contribution to problem difficulty, as the processing required in each case may draw on distinct cognitive functions. In Experiment 2 we explored these findings in more detail using a factorial design.

An experimental methodology has recently been employed by Beveridge, Jarrold, and Petit (2002) to examine the separability of executive functions. In a series of studies these authors independently manipulated the memory and inhibitory demands of three typical executive function tasks; the Stroop, the start-stop, and the continuous performance task. They argued that the presence of main effects of these factors, in the absence of an interaction, would suggest independent contributions of inhibition and memory to difficulty and consequently suggest memory and inhibition as dissociable executive functions. In contrast, an interaction might suggest that memory and inhibition are drawing on the same executive resource. We applied the same rationale to the current study. If the resolution gap and alternatives make independent contributions to difficulty then we would expect main effects of these variables. This would be consistent with an account that associates resolution with on-line planning or spatial memory and alternatives with inhibitory function. Alternatively an interaction would suggest that the two factors are drawing on the same resource: specifically an interaction that demonstrates a multiplicative relationship between alternatives and resolution.

EXPERIMENT 2

In order to explore the possibility that these factors are independent predictors of performance and to replicate Experiment 1, problems were taken from those given by Phillips et al. (1999)

and Gilhooly et al. (1999) that had crucial moves with a resolution gap of either one (low) or three (high) intermediate moves and had either one (low) or five (high) alternative moves. This enabled us to evaluate the extent that alternatives and resolution are independent contributors of difficulty.

Alternative moves were defined as follows. If there is only one alternative move then the moved disc could have only moved onto the alternative peg to that demonstrated. Under this category there can be no other legal moves possible either because of the way the discs are stacked (e.g., all the discs may be on the same peg) or because the remaining discs are in their final position. The definition includes an antilooping procedure such that it would be inefficient to move the same disc on two consecutive moves, and consequently the previously placed disc before the crucial move is not included as a viable alternative. Conversely, if there are five alternative moves the moved disc could move to the alternative peg but in addition there are two other discs that can be moved to either of the two pegs on which they are not currently placed. Examples of low and high alternative problems are given in Figure 2. The four conditions of Experiment 2 are therefore illustrated in Figures 1 and 2. Figure 1a is a low-resolution/low-alternative problem, and Figure 1b is a high-resolution/high-alternative problem. Figure 2a is a high-resolution/low-alternative problem, and Figure 2b is a low-resolution/high-alternative problem.

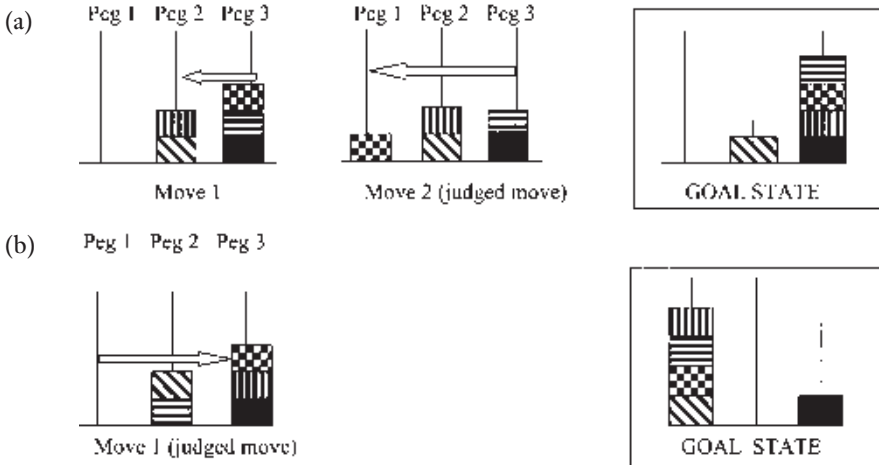


Figure 2. Illustration of the alternatives factor in the verification TOL. The arrow illustrates the movement of disc that the participant is asked to judge. In both examples a confirm move is illustrated, and the decision is required after the move has been completed. This move may occur in the middle of a solution. Therefore where appropriate the previous moves are shown. Figure 2a illustrates a low alternative problem; the vertical striped disc was moved in the previous move, and because an antilooping clause is included, it is not included as a possible alternative move. Discs in their final position are not included either. Therefore the only legal alternative moves are the movement of the checkered disc to Peg 2 or Peg 1. Figure 2b illustrates a high alternative problem. Here there are no previous moves or discs already in their goal position. Therefore there are six alternative moves because any of the three discs positioned uppermost of each peg can move to either of the two pegs on which they are not currently placed.

Method

Participants

A total of 26 undergraduates at the University of Plymouth participated for course credit. All had normal colour vision but none were familiar with TOL or had participated in Experiment 1.

Procedure

Participants were introduced to the TOL task as in Experiment 1 and were asked to verify or dispute a demonstrated move that should solve the problem in the minimum number of moves. They were asked to work as quickly as possible but avoid making errors where possible. One practice problem was issued, followed by 32 problems presented in a unique randomized order to each participant. These are given in Table 5. The experiment lasted approximately 30 minutes.

Tower of London task. The confirm decisions consisted of four problems with low resolution (one intermediate move) and low alternatives (one alternative move), four problems with high resolution (three intermediate moves) and high alternatives (five alternatives), four problems with low resolution and high alternatives, and four problems with high resolution and low alternatives. Matched dispute trials were produced using the procedure outlined in Experiment 1. The problems are detailed in Table 5.

Design

A 2 (confirm/dispute) \times 2 (high/low resolution) \times 2 (high/low alternatives) within-subjects design was employed. There were two dependent variables: errors and latency in milliseconds for correct trials.

Results

The raw data were treated in the same way as that in Experiment 1, with latency judgements greater than two standard deviations above the mean removed and replaced with the condition mean of all cells. Trimming did not alter the pattern of the results observed. The latency and the error data for Experiment 2 are presented in Table 4.

A 2 (decision) \times 2 (resolution) \times 2 (alternatives) ANOVA on the latency data showed no evidence for an effect of decision, $F(1, 25) = 2.302$, $MSE = 1,810,140$, $p = .142$, $\text{Eta}^2 = .084$. There was a main effect of resolution, $F(1, 25) = 6.516$, $MSE = 1,230,375$, $p = .017$, $\text{Eta}^2 = .207$, which replicated the effect of the manipulation in Experiment 1. There was also a main effect of alternatives, $F(1, 25) = 16.272$, $MSE = 1,022,234$, $p < .001$, $\text{Eta}^2 = .394$, showing that moves with high alternatives take significantly longer to evaluate than those with low alternatives, providing experimental confirmation of the results of the multiple regression in Experiment 1. Importantly, none of the interactions were significant: Decision \times Resolution, $F(1, 25) = 0.134$, $MSE = 1,692,842$, $p = .718$, $\text{Eta}^2 = .005$; Decision \times Alternatives, $F(1, 25) = 0.795$, $MSE = 1,553,910$, $p = .381$, $\text{Eta}^2 = .031$; Resolution \times Alternatives, $F(1, 25) = 1.05$, $MSE = 1,167,849$, $p = .315$, $\text{Eta}^2 = .040$; and Decision \times Resolution \times Alternatives, $F(1, 25) = 0.781$, $MSE = 1,130,523$, $p = .385$, $\text{Eta}^2 = .030$.

For the error data a 2 (decision) \times 2 (resolution) \times 2 (alternatives) ANOVA showed no evidence of an effect of decision type, $F(1, 25) = 1.144$, $MSE = 3.1$, $p = .295$, $\text{Eta}^2 = .044$. There

TABLE 4
 The effects of resolution and alternative move factors on response latency^a for correct trials and errors during confirm and dispute trials in Experiment 2^b

Resolution/ alternative	Confirm				Dispute			
	Response latency		Errors		Response latency		Errors	
	M	SD	M	SD	M	SD	M	SD
Low/low	2,643	1740	0.8	1.0	2,707	1678	1.8	1.1
Low/high	3,077	1448	1.7	1.2	3,711	1786	1.2	1.1
High/low	3,125	1628	1.1	1.0	3,318	1770	1.6	1.0
High/high	3,513	1788	1.4	1.1	3,754	1839	1.4	1.2

Note: On low resolution problems the disc is finally placed on the 2nd move. High-resolution problems are resolved on the 4th move. Low-alternative moves have one legal alternative move other than that demonstrated; high-alternative moves have five legal alternative moves other than that demonstrated.

^a In ms. ^b $n = 26$.

was no effect of resolution, $F(1, 25) = 0.008$, $MSE = 0.585$, $p = .928$, $\text{Eta}^2 < .001$, illustrating that three-resolution judgements were no harder than one-resolution judgements. There was no evidence of a difference between low- and high-alternative moves, $F(1, 25) = 1.26$, $MSE = 0.462$, $p = .272$, $\text{Eta}^2 = .048$, although there was a significant interaction between the decision type and the alternatives factor, $F(1, 25) = 11.29$, $MSe = 0.94$, $p < .01$, $\text{Eta}^2 = .311$. Planned comparisons on the significant interaction showed a significant effect of alternatives on confirm judgements, $F(1, 25) = 14.27$, $MSE = 0.57$, $p < .001$, with high-alternative judgements resulting in more errors ($M = 1.52$) than low-alternative judgements ($M = 0.96$). In contrast there was no effect of alternatives on dispute judgements, $F(1, 25) = 3.73$, $MSE = 0.84$, $p = .064$. Indeed with dispute judgements, the data show that low-alternative judgements result in more errors ($M = 1.67$) than high-alternative judgements ($M = 1.32$). This finding suggests that participants are more likely to dispute a move if there are a high number of alternative choices. This leads to more errors when alternatives are high on confirm judgements, but fewer errors on dispute judgements where this is the correct response.

There was a small three-way interaction between the decision, resolution, and alternatives factors, $F(1, 25) = 4.583$, $MSE = 0.765$, $p = .042$, $\text{Eta}^2 = .155$. This interaction is carried by an unusual pattern of errors within dispute trials, where low alternatives are associated with higher error rates than are high alternatives at low resolution. There is little difference as a function of alternatives at high resolution. This pattern contrasts with confirm trials where low alternatives are associated with fewer errors than high alternatives at both low resolution and, to a lesser extent, at high resolution. None of the other interactions were significant: Decision \times Resolution, $F(1, 25) = 0.006$, $MSE = 0.785$, $p = .938$, $\text{Eta}^2 = .000$, Resolution \times Alternatives, $F(1, 25) = 0.517$, $MSE = 0.456$, $p = .479$, $\text{Eta}^2 = .020$.

Discussion

The findings of Experiment 2 provide further evidence that the resolution gap and the number of alternative moves make independent contributions to the processing demands in

TABLE 5
Problems used in Experiment 2 and as Set A in Experiment 3

Problem no.	Trial type	Resolution/alternative	Start state			Moves demonstrated by computer	Goal state			
			Peg 1	Peg 2	Peg 3		Peg 1	Peg 2	Peg 3	
1	Confirm	Low/low		bpygr		b-3, p-3, y-3	pg	yr	b	
2				rgy	p	p-1, r-1		bp	yr	
3			bg	pr	y	y-2, b-2	py	r	bg	
4	High/low		g	bry	p	p-1, b-1	y	gp	br	
5					rgbyp	r-1, g-1	grb	py		
6				r	ygpb	y-2, g-1		r	pgyb	
7				ypgbr		y-1, p-1	rg	y	pb	
8				ybpgr		y-1	g	yr	pb	
9	Low/high		rp	by	g	g-2	yp		bgr	
10			p	rg	yb	p-3	ygpr		b	
11			prb	g	y	y-1	bp	r	yb	
12	High/high		yb	p	gr	y-2	gpy	rb		
13			y	grp	b	b-1		p	ybgr	
14			g	yrb	p	p-1		gb	pyr	
15			rbg	p	y	p-3	yp	bgr		
16			g	pr	by	b-2		rpbyg		
Dispute trials										
17	Dispute	Low/low	gyprb			g-2, y-2, p-3	pb	g	yr	
18			brp	y	g	y-3, b-2	gy	pbr		
19	High/low		yb	p	gr	p-1, g-2	b	gr	yp	
20			gbp	y	r	y-3, g-2	ry	gb	p	
21					brgpy		b-3, r-2	rbg	yp	
22				b	pryg		p-1, r-1		b	yrpg
23				pyrgb			p-3, y-2	p	yg	br
24	Low/high		pgyrb			p-2	rb	yg	r	
25			gp	r	by	r-3		grb	py	
26			r	pg	y	y-1		g	pryb	
27	High/high		r	p	ybg	p-1	b	pr	gy	
28			y	rb	pg	y-3	bg		ryp	
29			rby	g	p	g-1	y	pgrb		
30			pbg	y	r	y-1		rg	ypb	
31			y	p	bgr	y-3	grb		py	
32			yb	gp	r	g-3		bygpr		

Notes: Problems based on a selection given by Gilhooly et al. (1999) and Phillips et al. (1999). In the start and goal states discs are ordered so that the leftmost is uppermost. Problems 1–16 are confirm trials and show the correct move. Problems 17–32 are dispute trials. In dispute trials the move illustrated in the confirm trial is changed so that the same disc moves to the alternative but incorrect peg. Sets B and C were created by changing peg position and order and disc colour as described in the text. b = blue. r = red. y = yellow. g = green. p = pink. The moves are illustrated by giving the coloured disc that was moved and the peg that it was moved to. Low-resolution disc moves have one intermediate move before the final placing of the disc; high-resolution moves have three. Low-alternative problems have one legal alternative move other than that demonstrated; high-alternative move problems have five.

the verification task. The presence of main effects in the latency data and the absence of an interaction are consistent with an account that ties these factors to distinct cognitive operations. It is clear from these data that conflict moves cannot be viewed as a homogeneous group with a single defining characteristic. Instead they vary as a result of the degree of processing required, which, as we have shown, is a function of the number of moves that must be planned ahead and the number of competing alternative moves that there are to choose between.

Interestingly the key findings emerge in the latency data rather than the error data. Whilst there is some evidence that the number of alternative moves influences errors for confirm judgements, there is little evidence of increasing errors as a function of resolution gap. In general, the measure of primary concern in verification tasks is speed of response on correct trials. Errors are ordinarily viewed as a secondary indicator (Macleod, Hunt, & Mathews, 1978). The one notable effect in the error data in the present study was an interaction between type of judgement and number of alternatives. The interaction can be interpreted as resulting from a tendency to dispute a move when there are a high number of competing possibilities. This may suggest one reason why errors are less informative than latency data in this study. Errors may result from a number of sources, and in this case some participants are engaging in a form of strategic processing whereby there is a tendency to produce a particular response associated with specific task characteristics. Clearly such strategies will make interpretation of the error patterns difficult.

The evidence that the number of alternative moves contributes to problem difficulty is in some respects surprising. During both confirm and dispute trials the demonstrated move is the correct disc selection. In confirm trials it is moved to the correct peg and in dispute trials the incorrect peg. There is only one efficient solution, and hence the movement of a different disc is never correct. Despite the fact that alternative moves are never effective they nevertheless influence participants' response latency. It may be the case that alternatives are a less important predictor when participants are more practised at the task and learn that it is not necessary to consider all possible alternative moves.

In Experiments 1 and 2 participants were unfamiliar with the task, and it may be that the resolution gap and/or alternatives factors are a function of this novelty. Executive functioning is thought to be most important in novel situations, and it would be interesting to see how these characteristics affect performance as the participant becomes more practised. Consequently Experiment 3 explores the effect of the resolution gap and alternative moves as participants become more familiar with the task. This helps us assess whether the latency effects hold when error rates are reduced and the extent to which resolution and alternatives remain important contributors to difficulty as participants become more experienced.

EXPERIMENT 3

The aim of Experiment 3 was to examine the experimental factors amongst a group of participants as they gained experience on TOL. In order to improve performance the importance of accuracy was emphasized. Participants completed a block of 32 trials as novel subjects, so replicating Experiment 2, and then undertook some training trials in which they were given feedback about their performance and were asked to try to understand why they had got any problems wrong. They then completed a second set of 32 trials without feedback followed by a

second training session. Finally, they completed a final block of 32 problems without feedback as experienced participants.

In Experiment 1 we showed that SWM as assessed by TOL span was moderately negatively correlated with errors on TOL. In Experiment 3 we introduced a second SWM measure, Tic Tac Toe task (Daneman & Tardif, 1987) which has a heavy spatial processing requirement but does not share the superficial task demands of the TOL. Our aim was to examine the extent to which SWM continued to predict aspects of task performance amongst practised participants.

Method

Participants

A total of 47 undergraduates from the University of Plymouth participated for course credit. The data from three participants were dropped, one due to a computer malfunction and two because they were not administered all of the problems in the problem set.

Procedure

The task was introduced as in Experiments 1 and 2, but with the importance of reducing errors with practice emphasized. The tasks were administered in the same order to each participant. This order was: verification TOL 1, feedback version A, verification TOL 2, feedback version B, verification TOL 3. These were followed by the TOL span and the tic tac toe task. The experiment lasted about 1½ hours in total. The tasks used are described below.

Tower of London task. A computerized version of the five-disc, three-peg TOL (Ward & Allport, 1997) was used as in Experiments 1 and 2. Three sets of 32 problems were generated as given in Table 5. For each problem in Set A, matched versions for Sets B and C were created by changing peg orders and disc colours, so superficially changing the problem but maintaining the same logic. The TOL has three pegs, and by swapping the peg position there are six unique peg orders. Three of these were made confirm trials, and the remaining three had the disc moved to the alternative but incorrect peg, so producing matched dispute trials. In each of the six versions a different relationship between disc and colour was used.

The set used at each stage of familiarity was fully counterbalanced. Each set consisted of 16 confirm and 16 dispute judgements. There were four low-resolution/low-alternative judgements, four high-resolution/low-alternative judgements, four low-resolution/high-alternative judgements, and four high-resolution/high-alternative judgements. The problems within each set were administered in a unique random order to each participant. Instructions and scoring were as those given in Experiment 1.

Feedback version. The five-disc, three-peg (as Ward & Allport, 1997) computerized verification TOL task was used in Experiments 1 and 2 and was modified so that the participants were given accuracy feedback via an on-screen message about their response. All other program details were as previously described in the verification program, except that if the participant got a judgement wrong an on-screen button allowed them to replay the move as many times as they wished so they could work out why.

Two sets of 16 previously unused problems were used, 8 of which were confirm trials, and 8 were matched dispute trials. Problems were administered in the same order to each participant and were given in increasing difficulty as determined by their resolution gap. The results of this task were not formally analysed.

TOL span. A TOL span task was issued using the program details, scoring, and procedure given in Experiment 1.

Tic Tac Toe task. The procedure followed was as that in Daneman and Tardif's (1987) study. In a three-dimensional version of Tic Tac Toe (TTT), participants had to select the icons (noughts or crosses) that made a winning line on a grid. Each grid consisted of three planes (upper, middle, and lower). Each plane was divided into nine equal-sized squares. Each grid contained some red and blue tokens representing the pieces of the two players in the game. Embedded in the grid was one winning line—that is, three tokens of the same colour that formed a line that ran horizontally, vertically, or diagonally over a single plane or could be spread across three planes. The participant's task was to locate the winning line and highlight it with a mouse click. After two grids participants were given two 3-dimensional recall grids, one after the other, and they had to recall and reproduce the positions of the winning lines in the order in which they had been shown by using the mouse to click where the winning lines had appeared. Participants were given several practice items at the two-winning-lines level prior to the test starting. When the experiment proper started participants were warned to expect the winning lines per set to increase during the course of the test. There were three sets at Level 2, and then the level increased by one, up until three trials of five grids had been administered, resulting in a total of 12 sets (or 42 grids). A global scoring method was used in which 1 point was awarded for each winning line that the participant correctly recalled, provided it was recalled in the correct sequential position.

Design

A 2 (unpractised/practised participants) \times 2(confirm/dispute) \times 2 (high/low resolution) \times 2 (high/low alternatives) within-subjects design was employed. There were two dependent variables: errors and the latency in milliseconds for correct decisions. In addition two SWM measures were administered.

Results

Experimental analysis

The main interest of the present study was to explore the effect of the resolution gap and alternative moves on reaction times and errors when participants became increasingly familiar with TOL. Hence the analysis focuses on judgement latency and error rates for the problems presented in the first block and those presented in the final block of testing. The reaction time data showed some positive skewing and so were treated in the same way as those in Experiments 1 and 2 in accordance with Ratcliff (1993). This trimming did not alter the pattern of results that were observed. Descriptive statistics of the judgement, resolution, and alternatives factors are given in Table 6.

A 2 (familiarity) \times 2 (decision) \times 2 (resolution) \times 2 (alternatives) within-subjects ANOVA on the latency data showed evidence for an effect of familiarity, $F(1, 43) = 26.048$, $MSE = 8,641,374.086$, $p < .001$, $\eta^2 < .337$, illustrating that correct judgements were quicker when participants were practised than they were unpractised. There was no effect of decision, $F(1, 43) = 0.049$, $MSE = 3,438,629$, $p = .826$, $\eta^2 < .001$. There was a large main effect of resolution, $F(1, 43) = 34.004$, $MSE = 3,884,101.672$, $p < .001$, $\eta^2 = .442$, which replicated the findings of Experiments 1 and 2. There was also a large main effect of alternatives, $F(1, 43) = 122.772$, $MSE = 1,900,483.963$, $p < .001$, $\eta^2 = .741$, and this finding replicates the effect observed in Experiment 2. In contrast to Experiment 2, there was a significant interaction

TABLE 6
 The effects of the resolution gap and alternative move factors on response latency^a for correct responses and the number of errors for confirm and dispute trials in unpractised and practised participants in Experiment 3^b

Resolution/ alternative	Confirm				Dispute			
	Response latency		Errors		Response latency		Errors	
	M	SD	M	SD	M	SD	M	SD
	<i>Unpractised</i>							
Low/low	4,107	1,733	0.75	1.20	3,607	1,771	1.00	1.09
Low/high	5,106	2,525	1.30	1.09	5,720	2,166	1.25	1.18
High/low	5,235	3,332	1.30	1.09	5,271	3,235	1.25	1.01
High/high	5,507	2,306	1.23	1.08	5,489	2,190	1.13	1.17
	<i>Practised</i>							
Low/low	2,438	999	0.11	0.39	2,512	1,163	0.50	1.02
Low/high	4,129	2,028	0.73	0.69	4,434	1,880	0.68	0.88
High/low	3,837	1,780	0.73	0.76	3,807	1,968	1.27	0.76
High/high	4,935	2,060	0.35	0.61	4,803	1,902	0.45	0.73

^aIn ms. ^b*n* = 44.

between resolution and alternatives, which is illustrated in Figure 3, $F(1, 43) = 14.728$, $MSE = 3,357,874.429$, $p < .001$, $\eta^2 = .225$. This interaction did not reflect the multiplicative relationship between resolution and alternatives that one might expect if these factors are drawing on the same resource. Instead the interaction was carried by the presence of a very large effect of alternatives at low resolution, $F(1, 43) = 100.814$, $MSE = 2,468,000$, $p < .001$, and a smaller but still substantial effect of alternatives at high resolution, $F(1, 43) = 12.174$, $MSE = 2,790,359$, $p < .001$. There was an interaction between familiarity and alternatives, $F(1, 43) = 6.546$, $MSE = 1,690,472$, $p = .014$, $\eta^2 = .132$, illustrating a greater effect of alternatives in participants that were practised in TOL, $F(1, 43) = 98.638$, $MSE = 1,753,983$, $p < .001$, and a smaller but still substantial effect of alternatives for the unpractised subjects, $F(1,43) =$

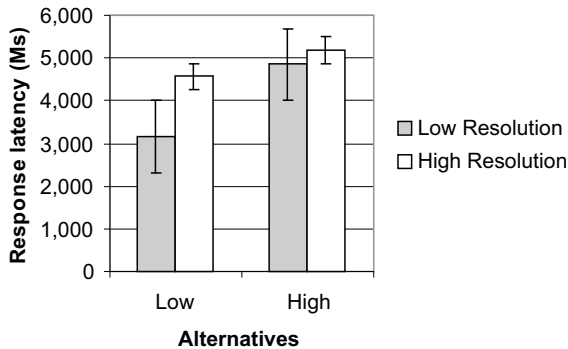


Figure 3. Resolution × Alternatives interaction in Experiment 3. (Response latency in milliseconds for correct trials.) Error bars indicate the standard error.

38.859, $MSE = 1,836,973$, $p < .001$. None of the remaining interactions were significant: Familiarity \times Decision, $F(1, 43) = 0.000$, $MSE = 3,767,973$, $p = .990$, $\text{Eta}^2 = .000$; Familiarity \times Resolution, $F(1, 43) = 0.793$, $MSE = 3,521,881$, $p = .378$, $\text{Eta}^2 = .018$; Familiarity \times Decision \times Resolution, $F(1, 43) = 0.342$, $MSE = 2,377,791$, $p = .562$, $\text{Eta}^2 = .008$; Familiarity \times Decision \times Alternatives, $F(1, 43) = 1.136$, $MSE = 1,610,310$, $p = .282$, $\text{Eta}^2 = .027$; Familiarity \times Resolution \times Alternatives, $F(1, 43) = 1.136$, $MSE = 2,435,519$, $p = .292$, $\text{Eta}^2 = .026$; Decision \times Resolution, $F(1, 43) = 0.575$, $MSE = 2,601,898.398$, $p = .452$, $\text{Eta}^2 = .013$; Decision \times Alternatives, $F(1, 43) = 1.665$, $MSE = 2,740,896.633$, $p = .204$, $\text{Eta}^2 = .037$; Decision \times Resolution \times Alternatives, $F(1, 43) = 3.246$, $MSE = 1,669,229.229$, $p = .079$, $\text{Eta}^2 = .070$; Familiarity \times Decision \times Resolution \times Alternatives, $F(1, 43) = 1.377$, $MSE = 1,753,573$, $p = .247$, $\text{Eta}^2 = .031$.

In the error data a 2 (familiarity) \times 2 (decision) \times 2 (resolution) \times 2 (alternatives) within-subjects ANOVA showed evidence of an effect of familiarity, $F(1, 43) = 32.723$, $MSE = 1.617$, $p < .001$, $\text{Eta}^2 = .432$, demonstrating that practised participants made more accurate judgements than unpractised participants. There was no effect of decision, $F(1, 43) = 2.183$, $MSE = 1.437$, $p = .147$, $\text{Eta}^2 = .048$. There was no effect of alternatives, $F(1, 43) = 0.188$, $MSE = 0.612$, $p = .667$, $\text{Eta}^2 = .004$, although the Decision \times Alternative interaction was significant, $F(1, 43) = 5.207$, $MSE = 0.766$, $p = .027$, $\text{Eta}^2 = .108$. The pattern of means for this interaction was similar to that in Experiment 2. Low-alternative moves resulted in fewer errors than high-alternative moves for confirm judgements (.72 vs. .90), $F(1, 43) = 5.184$, $MSE = 0.527$, $p = .029$, whereas for dispute judgements the opposite pattern was evident but the difference was not significant (1.0 vs. .88), $F(1, 43) = 1.614$, $MSE = 0.852$, $p = .210$. Once again this suggests that some participants are prone to dispute a move if there are a number of alternative possibilities. There was a main effect of resolution, $F(1, 43) = 5.732$, $MSE = 0.992$, $p = .021$, $\text{Eta}^2 = .118$, and a significant interaction between resolution and alternatives, $F(1, 43) = 38.031$, $MSE = 0.641$, $p < .001$, $\text{Eta}^2 = .469$. There was also a Familiarity \times Alternatives interaction, $F(1, 43) = 6.165$, $MSE = 0.467$, $p = .017$, $\text{Eta}^2 = .125$. Both of these two-way interactions were subsumed by a significant three-way interaction between familiarity, resolution, and alternatives, $F(1, 43) = 9.470$, $MSE = 0.304$, $p = .004$, $\text{Eta}^2 = .180$, illustrating that in the unpractised group there was an effect of alternatives at low resolution, $F(1, 43) = 11.834$, $MSE = 0.588$, $p = .001$, but not at high resolution, $F(1, 43) = 0.546$, $MSE = 0.666$, $p = .463$. Amongst practised participants there was also a significant effect of alternatives at low resolution, $F(1, 43) = 20.236$, $MSE = 0.344$, $p < .001$, but surprisingly at high resolution more errors were made with low alternatives than with high alternatives, (1.0 vs. .40), $F(1, 43) = 37.523$, $MSE = 0.425$, $p < .001$. We consider this interaction in more detail in the Discussion section that follows. None of the remaining interactions were significant: Familiarity \times Decision, $F(1, 43) = 1.754$, $MSE = 1.362$, $p = .192$, $\text{Eta}^2 = .039$; Familiarity \times Resolution, $F(1, 43) = 116$, $MSE = 0.602$, $p = .735$, $\text{Eta}^2 = .003$; Familiarity \times Decision \times Resolution, $F(1, 43) = 1.957$, $MSE = 0.610$, $p = .169$, $\text{Eta}^2 = .044$; Familiarity \times Decision \times Alternatives, $F(1, 43) = 1.336$, $MSE = 0.562$, $p = .254$, $\text{Eta}^2 = .030$; Decision \times Resolution, $F(1, 43) = 0.003$, $MSE = 0.464$, $p = .956$, $\text{Eta}^2 = .000$; Decision \times Resolution \times Alternatives interaction, $F(1, 43) = 0.261$, $MSE = 0.657$, $p = .612$, $\text{Eta}^2 = .006$; Familiarity \times Decision \times Resolution \times Alternatives, $F(1, 43) = 0.394$, $MSE = 0.436$, $p = .534$, $\text{Eta}^2 = .009$.

Correlational analysis

The reaction time and errors for odd trials and even trials were determined for all individual difference measures, and the reliability of the measures was established with Cronbach's alpha statistic. In each case this was calculated by analysing the problems in the order in which they were administered, with the exception of the TOL verification paradigm where problems were first grouped by condition before being sorted in order of presentation. The statistics are determined as $\alpha = .84$ for error judgements and $\alpha = .96$ on latency judgements for the verification paradigm. The TOL span accuracy yields an alpha of .84 and the TTT span errors an alpha of .90.

TOL errors were examined in unpractised and practised performance to examine the role of SWM as participants became more experienced with the task. There was no relationship between TOL span and the latency measures on the verification TOL. However, correlations emerged between error rates and measures of SWM. There was a significant correlation between errors and the TOL span amongst both unpractised ($r = -.454, p = .002$) and practised participants ($r = -.319, p = .037$). Although there was no correlation with between errors and TTT span amongst unpractised participants ($r = -.150, p = .331$), a significant correlation emerged in practised participants ($r = -.326, p = .031$). Although the correlation between errors and TOL span is stronger for unpractised than for practised participants, the opposite pattern is present for the TTT span. Consequently the data provide no evidence that experience on the task reduces the demands placed upon executive processes.

Discussion

The main purpose of Experiment 3 was to examine the effects of resolution and alternatives amongst participants as they gained familiarity with TOL. The training manipulation significantly reduced error rates, and yet the main effects of Experiment 2 were replicated. In contrast to Experiment 2, in Experiment 3 an interaction between the resolution and alternatives factors emerged. This could indicate that the effects of resolution and alternatives can be explained by the same underlying process. However, this interpretation depends upon an interaction in which the effects of alternatives are exacerbated by high resolution, and vice versa. The interaction is carried by a completely contrary pattern, where the effects of alternatives, whilst substantial for both low and high resolution, are of lower magnitude at high resolution. Hence the findings in the latency data remain consistent with the view that the two factors are drawing on distinct resources.

The pattern of findings in the error data are more difficult to interpret. The alternatives by decision interaction replicates the finding of Experiment 2, but the interaction between resolution, alternatives, and familiarity is unique to Experiment 3. This interaction reflects more errors for high-alternative problems than for low-alternative problems at low resolution for both unpractised and practised participants. There is no effect of alternatives at high resolution for unpractised participants, but surprisingly at high resolution amongst practised participants high alternatives are associated with fewer errors than are low alternatives. A number of potential explanations for the interaction can be discounted. It is not a peculiarity of the problems used in Experiment 3, as these were structurally identical to those used in Experiment 2. The interaction cannot be explained in terms of a speed-error trade-off since there is no evidence of a complementary effect in the latency data, and there is no correlation between

the latency and error measures ($r = -.076$, $p = .524$). Indeed the contrast between the interactions in the error and latency data makes it particularly difficult to provide a consistent explanation of both effects. It is important to emphasize, however, that the error rates in Experiment 3 are, particularly in practised participants, low. Consequently any effects present need to be interpreted with some caution, and additional research is needed in order to examine whether these findings are replicable.

Despite the anomalous finding in the error data the results once again indicate that resolution and alternatives are predictors of move latency on the verification task. These effects remain robust even amongst participants who have experienced many trials. This fact raises certain issues that require further discussion, given that the executive is thought to be evoked during novel situations (e.g., Baddeley & Logie, 1999). It may be possible to argue that these factors are not executive in nature; however, if this were true then there must be other performance characteristics that are, but it is almost impossible to imagine what other factors there could be that are independent of resolution evaluation and alternative-move assessment. We would argue that even though participants are familiar with the task, each problem still has to be considered on its individual characteristics, and it is unlikely that participants can ever become so practised that they are able to automatically evaluate TOL moves, such as described by Norman and Shallice's (1986) contention scheduling system. However, practice may speed up such an assessment, and once the problem has been assessed discs may be moved in an automated way (if the standard TOL paradigm were employed).

The individual differences analysis once again illustrated the role of SWM in the TOL verification task. The passive TOL memory span was moderately correlated with errors on the task, such that higher span scores are related to fewer errors on TOL. This relationship exists in both unpractised and practised participants. This supports previous findings (e.g., Phillips et al., 1999) and suggests a spatial component to TOL. The evidence that SWM remains predictive of performance even amongst participants that have become practised on the task provides support for the claim that even practised participants draw upon executive resources in solving the verification version of the TOL.

GENERAL DISCUSSION

The TOL has been widely used in both theoretical and clinical psychology, yet little is known about the performance requirements or what the task actually measures. We have shown that an analysis of performance at the level of the individual move avoids some of the problems traditionally associated with the "planning" explanation of TOL, but is not inconsistent with this research. In this task no time for planning is given prior to witnessing the moves, and the concepts of resolution and alternative-move assessment are likely candidates for the mental activity that takes place during on-line planning. Indeed Ward and Allport have implicated concepts similar to these in their Move Selection Framework (Ward & Allport, 1997), which suggests a procedure for selecting moves in TOL.

Given the reported importance of the conflict move in TOL performance these three experiments have demonstrated that the notion of a conflict move is not straightforward and that they are not a homogenous group. In fact there are a range of factors contributing to their difficulty, including the number of moves until a disc is resolved and the quantity of legal alternative moves that are possible. These characteristics predict a substantial

proportion of the variance in latency data and a good proportion of the error variance. The advantage of our approach has been that we have been able to provide clear process-based explanations for the difficulty of conflict problems, which contrast with traditional but inexact characterizations based upon the notion that difficulty is a result of a judgement that is counterintuitive.

Frontal patients have particular problems with conflict moves (Goel & Grafman, 1995; Morris et al., 1997), and normal participants often make their first mistake at the first conflict move (Miyake et al., 2000b), suggesting that the conflict move is a key component in TOL performance. Having examined the performance requirements of TOL in some depth we argue that one reason why conflict moves are particularly difficult is that the participant must track the outcome of the proposed move until the disc is finally placed. The more intermediate moves there are the more difficult this is. Following this line of argument, forward moves are easier than conflict moves because the resolution is immediate. The tracking of a disk to its goal position can be conceived of as a form of on-line planning and is likely to draw upon the resources of WM. Indeed the data presented here and elsewhere (e.g., Phillips et al., 1999) suggest that SWM capacity is a good predictor of performance on the task. However, whilst the resolution gap is a source of difficulty it is probably not the source of the performance deficit noted in frontal patients as their SWM is not generally impaired (Goel & Grafman, 1995; Owen et al., 1990).

A second factor that affects move difficulty is the number of alternative moves that could be made. This factor applies to both forward and conflict moves. Previous research into alternatives in TOL has examined points in the solution where there are alternative valid solutions and how these may impact on difficulty (e.g., Ward & Allport, 1997). However, our studies provide evidence that alternative moves contribute to difficulty, even when there is only one valid solution. This is particularly true when the resolution is low (one intermediate disc move before resolution), and furthermore it continues to be the case as task familiarity increases. This argument should apply at any point in the standard version of TOL where the participant evaluates a move that they are considering. The ability to discriminate between competing alternatives and resist inappropriate choices in favour of more appropriate ones is commonly associated with executive control and is likely to have an inhibitory component. It is possible that the manipulation of alternatives draws on inhibitory function, whereas the manipulation of resolution gap draws upon spatial WM resources. Hence the TOL deficit in frontal patients may be a result, not from the SWM demands of conflict problems, but instead from the inhibitory demands resulting from deciding between competing move choices. This argument is consistent with recent claims that performance on tower tasks is related to inhibition (Miyake et al., 2000a; Welsh et al., 1999).

The evidence presented in this paper is consistent with the notion that resolution and alternatives are associated with distinct cognitive operations. This is particularly the case with the latency data, where in Experiment 1 resolution and alternatives emerged as independent predictors of performance in the multiple regression analysis. In Experiment 2 there were main effects of both factors in the latency data, but no interaction. In Experiment 3 these main effects were replicated, and although an interaction between the factors emerged, this reflected a smaller (although still substantial) effect of alternatives at high resolution than at low resolution. If both factors were drawing on the same resource the

opposite pattern should have been observed. If the characterization of these factors as drawing on planning/SWM and inhibition is correct then it is possible to reconcile the conflicting claims made concerning the cognitive processes involved in the TOL (e.g., Goel & Grafman, 1995; Phillips et al., 1999; Ward & Allport, 1997). It may indeed be the case that the task draws upon some form of planning function, in the form of the on-line planning that is involved in considering the consequences of a proposed move. It may also be the case that the task requires, at certain points, the inhibition of some alternative move(s) in favour of others. Both of these factors contribute to the overall difficulty of a given problem. Whilst the data presented here cannot be taken as firm evidence for this position, as the claim relies on what are essentially null effects, the data do provide evidence against the view that the TOL is measuring a single cognitive construct.

The detailed analysis of the TOL presented here has been achieved by use of a previously unused verification paradigm. This series of experiments suggests that this paradigm has several advantages over the standard version of TOL, where the participant makes the moves from start to goal themselves. Notably the task allows performance to be analysed at the level of the individual move, and this avoids the difficulties inherent in interpreting global measures of performance over a full problem set. However, a difficulty with the verification approach is that one cannot gain direct insight into an individual's preferred move choice at that point in the solution path. In addition, although the latency data have been relatively clean throughout the experimental series, the error data have been difficult to interpret at times, and a question remains as to whether latency or errors are the best measure of processing requirements.

An alternative approach would have been to use a paradigm in which the participant selects and makes the crucial move. This is certainly a consideration for future research, as it would add richness to the data because preferred move choices would be elicited. However, the verification task has a key advantage over either a partial or a full participant-controlled disc-moving task. In the verification paradigm decision time is unaffected by the distance that the disc must be moved and is consequently likely to be less influenced by motor speed. This represents an important advantage particularly if the task was administered to clinical populations.

In addition, recent evidence suggests that the factors identified with the verification task can be applied in making predictions concerning errors and latency for the completion of a full TOL task. Carder, Perfect, and Handley (2003) manipulated the overall resolution and alternative move demands in a set of six-move problems using a traditional TOL paradigm and administration. The findings of this study demonstrated that resolution and alternative moves remain predictors of performance, demonstrating that the verification paradigm data provides a valid exploration of TOL performance demands. Whilst it remains to be seen whether the verification TOL can be used successfully within a clinical setting, the findings presented here provide a valuable indication of the specific cognitive contributors to the difficulty of the task.

REFERENCES

- Baddeley, A. D. (1998). The central executive: A concept and some misconceptions. *Journal of the International Neuropsychological Society*, 4, 523–526.

- Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 28–61). Cambridge, UK: Cambridge University Press.
- Beveridge, M. J., Jarrold, C., & Petit, E. (2002). An experimental approach to genetic fingerprinting in young children. *Infant and Child Development*, *11*, 107–123.
- Carder, H. P., Perfect, T. J., & Handley, S. J. (2003). *Ageing and Tower of London: Alternative moves, resolution gap and executive function as predictors of task performance*. Poster presented at the BPS Workshop on the Neuropsychology of Ageing, Holland House, UK.
- Daneman, M., & Tardif, M. (1987). Working memory and reading skill re-examined. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 491–508). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Gilhooly, K. J., Phillips, L. H., Wynn, V., Logie, R. H., & Della Sala, S. (1999). Planning processes and age in the five-disc Tower of London task. *Thinking and Reasoning*, *5*, 339–361.
- Gilhooly, K. J., Wynn, V., Phillips, L. H., Logie, R. 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*, 165–178.
- Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in “planning” functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia*, *33*, 623–642.
- Handley, S. J., Capon, A., Copp, C., & Harper, C. (2002). Conditional reasoning and the Tower of Hanoi: The role of spatial and verbal working memory. *British Journal of Psychology*, *93*, 501–518.
- Macleod, C. M., Hunt, E. B., & Mathews, N. N. (1978). Individual differences in the verification of sentence–picture relationships. *Journal of Verbal Learning and Verbal Behaviour*, *17*, 493–507.
- Miyake, A., Emerson, M. J., & Friedman, N. P. (2000a). Assessment of executive functions in clinical settings: Problems and recommendations. *Seminars in Speech and Language*, *21*, 169–183.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000b). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.
- Morris, R. G., Miotto, E. C., Feigenbaum, J. D., Bullock, P., & Polkey, C. E. (1997). Planning ability after frontal and temporal lobe lesions in humans: The effects of selection equivocation and working memory load. *Cognitive Neuropsychology*, *14*, 1007–1027.
- Norman, D. A., & Shallice, T. (1986). Attention to action. Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol 4. pp. 1–18). New York: Plenum Press.
- Owen, A. M., Downes, J. D., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*, 1021–1034.
- Owen, A. M., Doyon, J., Petrides, M., & Evans, A. C. (1996). Planning and spatial working memory: A positron emission tomography study in humans. *European Journal of Neuroscience*, *8*, 353–364.
- Parker, D. M., & Crawford, J. R. (1992). Assessment of frontal lobe dysfunction. In J. Crawford, D. Parker, & W. Mckinlay (Eds), *A handbook of neuropsychological assessment* (pp. 267–291). Hove, UK: Lawrence Erlbaum Associates Ltd.
- Parkin, A. J. (1998). The central executive does not exist. *Journal of the International Neuropsychological Society*, *4*, 518–522.
- Phillips, L. H., Wynn, V., Gilhooly, K. J., Della Sala, S., & Logie, R. H. (1999). The role of memory in the Tower of London task. *Memory*, *7*, 209–231.
- Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London. *Quarterly Journal of Experimental Psychology*, *54A*, 579–597.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.
- Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London*, *B*, *298*, 199–209.
- Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.
- Shallice, T., & Burgess, P. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*, 727–741.

- Shnirman, G. M., Welsh, M. C., & Retzlaff, P. D. (1998). Development of the Tower of London-revised. *Assessment*, 5, 355–360.
- Ward, G., & Allport, A. (1997). Planning and problem solving using the five-disc Tower of London task. *The Quarterly Journal of Experimental Psychology*, 50A, 49–78.
- Welsh, M. C., Satterlee-Cartmell, T., & Stine, M. (1999). Towers of Hanoi and London: Contribution of working memory and inhibition to performance. *Brain and Cognition*, 41, 231–242.

Original manuscript received 18 March 2003
Accepted revision received 29 September 2003
PrEview proof published online date /month /year

Copyright of Quarterly Journal of Experimental Psychology: Section A is the property of Psychology Press (T&F) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.