Cognitive Engineering Laboratory

Practical Problem Solving in a Design Microworld: An Exploratory Study

Klaus Christoffersen

CEL 96-03

Cognitive Engineering Laboratory University of Toronto Department of Mechanical & Industrial Engineering
5 King's College Rd. Toronto, Ontario, Canada M5S 3G8
Phone: (416) 978-7399 Fax: (416) 978-3453
Email: benfica@mie.utoronto.ca URL: http://www.ie.utoronto.ca/IE/HF/CEL/homepage.html

Cognitive Engineering Laboratory

Director: Kim J. Vicente, B.A.Sc., M.S., Ph.D.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Mechanical & Industrial Engineering, and is one of three laboratories that comprise the U of T Human Factors Research Group. CEL began in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments, with a particular emphasis on power plant control rooms. Professor Vicente's areas of expertise include advanced interface design principles, the study of expertise, and cognitive work analysis. Thus, the general mission of CEL is to conduct principled investigations of the impact of information technology on human work so as to develop research findings that are both relevant and useful to industries in which such issues arise.

Current CEL Research Topics

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Defense and Civil Institute for Environmental Medicine, Japan Atomic Energy Research Institute, Natural Sciences and Engineering Research Council of Canada, Rotoflex International, and Westinghouse Science & Technology Center. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Current CEL projects include:

- Studying the interaction between interface design and adaptation in process control systems.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing a better understanding of the design process so that human factors guidance can be presented in a way that will be effectively used by designers.
- Designing novel computer interfaces to display the status of aircraft engineering systems.
- Developing and evaluating advanced user interfaces (in particular, transparent UI tools) for 3-D modelling, animation and painting systems.

CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente at the address printed on the front of this technical report.

ABSTRACT

The research presented here represents an attempt to bring the study of design problem solving into the laboratory. The intent of the study was explicitly exploratory; the goal was not to provide answers so much as it was to generate useful and interesting questions, and to serve as a springboard for future research. A novel task domain was proposed for use as a "design microworld". This environment was evaluated against commonly accepted characteristics of design problems and found to share many important features with such problems. An experiment was conducted to attempt to uncover some of the characteristics of subjects' problem solving strategies on these problems. Subjects of high and low spatial ability were asked to perform trials under various conditions related to the level of constraint in the problems. The results revealed that spatial ability and constraint level both have significant effects on subjects' performance. A speculative model was proposed as a potential mechanism through which spatial ability and constraint level may have combined to lead to the results observed. This model and many other interesting phenomena observed during the study were used in forming a number of suggestions of potentially fruitful avenues for continued research.

ACKNOWLEDGMENTS

Funding for this study was provided by research and equipment grants from the Natural Sciences and Engineering Research Council of Canada.

This document and the study it describes would not have been possible without the help, cooperation, and patience of a large number of people, to all of whom I offer sincere thanks. To Tim Yee Lee and Renée Chow who made significant contributions to the initial stages of data gathering and analysis. To the graduate students of the Cognitive Engineering Laboratory at the University of Toronto for their enthusiasm and many helpful comments. To Phil Smith for generously agreeing to let me commandeer his video equipment for the bulk of the protocol analysis. To Dave Woods for his support and patience. To Emily Chapman, in part for her help in conducting the experiment, but infinitely more for her absolutely steadfast support, encouragement, and understanding. Finally to Kim Vicente for his innumerable intellectual contributions, guidance, and for encouraging excursions such as this one down roads less traveled.

K.C.
July, 1996

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. PRELUDE

On April 11, 1970, NASA launched the 13th Apollo mission, the 3rd moon landing attempt, less than a year after Apollo 11's triumphant touchdown on the moon's surface. Fifty-five hours into the flight, and nearly 200,000 miles from earth, an explosion in one of the ship's oxygen tanks caused the power and oxygen supplies to begin to rapidly drain from the command module. Facing a four day journey back to earth, the three-man crew was forced to shut down the command module and retreat to the lunar module, a craft designed to support two people for a maximum of two days. Among a myriad of other problems, the crew faced the danger of being asphyxiated by their own carbon dioxide waste gases. The lunar module was equipped with an air scrubbing system, which used lithium hydroxide cartridges to trap CO_2 molecules and filter them from the air. However, the supply of lithium hydroxide in the lunar module would become saturated in less than thirty-six hours. Although the command module was equipped with a similar air scrubbing system which had more than enough lithium hydroxide for the trip, there was a critical problem. The lithium hydroxide cartridges from the command module were square-shaped; the receptacles in the lunar module were round. While the astronauts dealt with various other concerns, the members of the crew systems division at NASA were faced with the problem, almost literally, of trying to fit a square peg into a round hole.

Given this goal, the limited supply of non-essential materials on board the spacecraft, and the constraints represented by the parts of the air scrubbing system already in place in the lunar module, the crew systems team needed to design a system which could somehow bridge the gap between the lithium hydroxide cartridges from the command module and the receptacles in the lunar module. And they needed to do it quickly...

2. INTRODUCTION

While the scenario described in the prelude is certainly exotic, the characteristics of the problem which confronted the crew systems team at NASA that day in 1970 do not seem to be uncommon at all. The problem consists of a well defined goal, a finite set of materials, each with a limited set of possible functions, and a set of constraints due to the pre-existing structure of the specific environment. The task is to employ the materials in such a way that they somehow bridge the functional gaps in the existing structure, so that the resulting system of components accomplishes a specific goal. The goal is not to devise an optimal or elegant solution, but simply to design a system, using whatever means available, that serves to achieve the desired function.

The claim is made that these characteristics provide a coarse definition for a class of problems with many similarities to characteristics normally associated with design, thus distinguishing them from the types of problems most typically examined in contemporary problem solving research. However, they also differ from some of the commonly accepted features of design problems in ways that make them amenable to controlled study. Such problems therefore seem particularly well-suited as a starting point for research which seeks to examine factors associated with problem solving performance in this and similar classes of design-like scenarios.

The study presented here employs a novel experimental task context to attempt to explore some of the factors influencing subjects' performance in solving such problems. The study is essentially exploratory in nature. The goal is to reveal aspects of the domain and subjects' performance which serve to inspire and suggest directions for continued research. The study is in the spirit of what Hunt (1991) has termed an "engineering investigation", in which the goal is to uncover characteristics of subjects' heuristics and strategies that may eventually prove useful as guides towards methods for aiding practitioners confronted by similar problems. The goal might therefore be said to make steps towards the formation of principles useful for cognitive engineering (Woods and Roth, 1988) in environments, such as design, where problems of the sort described above are encountered.

In the following chapter, a brief discussion of some of the various perspectives to emerge in modern problem solving research will be presented to help put the current study in context. This will be followed by a review of some of the commonly accepted features of design problems. The experimental domain for the study will then be presented and evaluated against the proposed characteristics of design.

3. BACKGROUND

3.1 PREVIOUS RESEARCH

Hunt (1991) has distinguished three types of investigations within problem solving research: scientific investigations, engineering investigations, and humanistic investigations. The differences lie primarily in the theoretical goals set by researchers and the methods and conclusions which those goals allow. In scientific investigations, the goal is typically to form simple, yet general principles through structured empirical observation. This normally requires tightly controlled experimental situations in closed environments with few variables. To be accepted as scientific explanations, the derived principles must usually be general enough to unify results from various settings and must be specified precisely enough to be expressed formally. To take the opposite extreme, humanistic investigations are characterized by largely descriptive case studies of very complex problem solving processes involving open systems with often unenumerable sets of variables, taking place over extended periods of time. The goals are often to provide demonstrations which make important aspects of particular cases apparent, and to attempt to identify critical decisions or events. The results are rarely objective, and debates about the conclusions are common. Finally, engineering investigations are those which seek to derive knowledge which can be usefully applied to solving the types of problems under study. The systems studied are often open, but the studies are typically designed to minimize the influence of variables which are not explicitly accounted for. The goal is to describe behaviour at a practical level, i.e., one from which direction can be drawn regarding the design of tools or instructions which will aid practitioners in solving the sorts of problems examined.

3.1.1 Scientific Investigations

All three of Hunt's classes of investigation are easily identifiable within problem solving research, but the most common class of investigations is clearly the scientific class. Modern problem solving research has been dominated by the information processing (IP) viewpoint, as exemplified by the seminal work of Newell and Simon (1972). The goals of their studies were avowedly scientific and reductionist. They sought to identify the basic information processes by which humans solved problems, and to do so precisely enough so that computer simulations could be constructed which accurately reproduced the (successful forms of) behaviour they observed. Partly because of the level of detail necessary to describe behaviour in such terms, and partly because the artificial intelligence

community had already studied them, the problems they chose to use in their experiments were highly circumscribed, well-defined, formal tasks such as cryptarithmetic and symbolic logic.

Newell and Simon's characterization of problem solving as a basic two-step process of representation and heuristic search has proven extremely robust and has served as a basis for a great deal of subsequent research. Despite the success which the results of Newell and Simon have demonstrated as scientific descriptions of certain types of problem solving, their work and work like it has nonetheless been criticized on the basis of the problems that they chose to study. For instance, Rasmussen (1985) has argued that because these types of problems represent formal systems, which are defined only at a single level of abstraction, the analysis of such problems may ignore holistic aspects of reasoning associated with more complex, less artificial problems. Also, Wertheimer (1985) has criticized the computationally-oriented IP viewpoint in general on the basis that it does not address the fundamental processes which result in the understanding of what is crucial about a given problem and why. In any case, given the scientific orientation of work such as Newell and Simon's, there may be reason to question the use of the results as an immediate basis for engineering explanations of behaviour in more practical domains.

That the basic mechanisms identified by Newell and Simon are not sufficient for describing problem solving behaviour in more complex domains has been amply demonstrated by the recent work in the area of expertise and schema-driven problem solving. Compared to the tasks posed to their subjects by Newell and Simon, these domains require large amounts of specialized knowledge on the part of subjects. The task domains studied have included physics, bridge, chess, computer programming, and others (see, e.g., Chi, Glaser, and Farr, 1988).

The most salient results from this line of research have been those surrounding experts' use of schemas as a way to organize and embellish problem representations by activating associated relevant knowledge, and as a way to provide guidance in solution procedures (VanLehn, 1989). The recognition that access to large amounts of domain-specific knowledge plays a key role in skilled problem solving has been a major extension to Newell and Simon's work. The word extension is used here because many of the studies within this type of research, although concerned with more complex problems, have remained firmly rooted in the IP viewpoint, focusing primarily on scientific investigations. The goals of these studies have most often been to advance theories of expertise in general, and have been only secondarily, if at all, concerned with aiding practitioners in the domains under study. Producing computational models of observed behaviour has continued as a major goal of much of the research, serving to test the

sufficiency of the hypotheses produced. Ericsson and Simon (1993), for example, have stressed the view that the rigor of a problem solving theory is associated with the ability to provide rules which can form the basis for detailed simulations of the problem solving process. Thus, many of the problems studied have, in spite of their complexity, remained relatively well-defined, with little or no ambiguity surrounding the definition of states, operators, or solutions.

3.1.2 Humanistic Investigations

In stark contrast to these types of problems, recent work on "ill-structured" problems has attempted to study problem solving behaviour in domains as complex as national policy formation (e.g., Voss and Post, 1988; Voss, Wolfe, Lawrence, and Engle, 1991). These studies, for the most part, represent humanistic investigations in that the problem domains have constituted highly open systems with ill-defined states, operators, and solution paths. The solutions generated by subjects have been open to considerable debate, both in terms of the methods used, and the appropriateness of the final solutions themselves (this may be contrasted with problems in symbolic logic or basic physics, where there is rarely any room for interpretation of the correctness of a solution). Voss and Post (1988), in a summary of several investigations of problem solving in ill-structured environments, have acknowledged that only very general conclusions have been reached thus far, and that there are many gaps in the current understanding of problem solving in such domains. There has been relatively little progress in accounting in detail for the wide range of solution methods that exists within subjects, not to mention between subjects. Nonetheless, these humanistic studies, through a slow process of corpus building, are serving to expose and clarify some of the major characteristics of problem solving in these domains.

In an attempt to study similar problems from a more scientifically oriented viewpoint, there has also been a good deal of recent work on so called "complex problem solving" (see Frensch and Funke, 1995 for an overview). These studies have employed very complex microworld simulations, sometimes with thousands of interconnected variables. Despite the control gained by moving the study of complex problem solving into laboratory settings, the scale of the complexity has led in some cases to the same problem as encountered in other research on ill-structured problems. That is, evaluation of subjects' performance becomes difficult and subjective, and some of the conclusions thus take on a humanistic character. Also, many of the studies have attempted to study "complex problems" per se, and therefore the tasks have often been somewhat artificial, with

relatively little consideration given to the domains to which the results are intended to generalize, if any.

3.1.3 Engineering Investigations

Engineering investigations into problem solving can perhaps be most easily identified within the cognitive engineering community. The studies vary widely in methodology, ranging from ethnographic studies to relatively controlled laboratory experiments using microworld simulations. They also vary widely across domains, including process control, aviation, medical informatics, and others. Despite this variability and the often extensive use of results from scientific investigations (e.g., from within the cognitive psychology literature), the goals of such studies are relatively uniform in the desire to understand and eventually improve the ability of practitioners to solve problems in these specific environments. The intent is thus to be prescriptive as well as descriptive. It is these goals which mark them as prototypical engineering investigations.

3.1.4 Studies of Design

Studies in design methodology have tended to embody elements of both humanistic and engineering investigations. The focus has typically been on observational case studies within the dominant domains of design activity such as architecture, planning, engineering design, and software design, among others. While these investigations have necessarily had a humanistic flavour, the goals of the studies have often resembled those of engineering investigations. Given that design is an inherently applied discipline, one of the primary goals of investigations into design is often to help designers perform their jobs more thoroughly and efficiently (see e.g., Thomas and Carroll, 1984). Therefore, many studies specifically concerned with design and strategies for design attempt to draw implications for the practice of design in naturalistic settings (e.g., Darke, 1984; Akin, 1984, Lawson, 1984; Malhotra, Thomas, Carroll, and Miller 1980; Carroll, Thomas, and Malhotra, 1979).

The current study represents an attempt to study a form of design problem solving, as represented by the experimental task, from a cognitive engineering perspective. In this sense, the study is intended, in Hunt's terms, as an engineering investigation, with the goal of eventually leading towards improved aids for practitioners. The task used is meant to represent some of the significant characteristics of design problems but is nonetheless suitable for use in a relatively controlled experimental setting, as contrasted with many previous studies of design problem solving (e.g., Darke, 1984). Therefore, the study represents a compromise between the highly simplistic tasks of some of the early scientific

investigations of problem solving and the quasi-engineering investigations of some design researchers.

3.2 CHARACTERISTICS OF DESIGN PROBLEMS

In the most general terms, design can be said to be the process of intentionally changing an existing situation into a preferred one (Simon, 1969). Simon contrasts design with endeavours in the natural sciences which are concerned with discovering how things are, as opposed to deciding how things ought to be. Design then, is a goal-directed activity, the product of which is associated with some more or less specific purpose. Many more concrete definitions of design have been attempted, but each is typically tailored to a certain model or viewpoint (see Burns and Vicente, 1994 for a number of examples) and as such does not seem suitable as a general definition. As Goel and Pirolli (1992) have pointed out, the term "design" has been used to describe such a wide range of activities that it is often in danger of losing any real meaning. As one might expect then, there is no generally agreed upon way of identifying a "design" problem as distinct from a "non-design" problem.

How then, to characterize design problems? The term "problem" will be assumed to correspond to Duncker's (1945) definition: "A problem arises when a living creature has a goal but does not know how to reach it" (p.1). That is, there is a desired situation which is different from the current situation and cannot be reached in a direct, obvious way. This is stated merely to emphasize the notion of problem solving as a goal-directed, purposeful activity which is not carried out by simply retrieving a solution from memory. So what distinguishes design problems from other problems? In spite of the lack of an accepted definition, there are certain features which are cited repeatedly when discussing what identifies design problems. These include the following:

1. Design problems are inherently ill-structured (also referred to as ill-defined). The objective and/or the available means for achieving the objective are often initially poorly defined (Logan and Smithers, 1993; Rittel and Webber, 1973; Goel and Pirolli, 1992; Carroll, Thomas, and Malhotra, 1979; Schon, 1988; Vicente, Burns, and Pawlak, 1993). This is often noted as the most characteristic feature of design problems and can have many implications for the ways in which design problems are solved. Because this is such a dominant feature of design problems, a more detailed discussion of what it means for a problem to be ill-structured will be presented.

A problem can be ill structured with respect to any, or all of, the goal, the materials (the "givens") available for constructing a solution, and the available ways of using the givens to construct a solution. Voss and Post (1988) outline Reitman's description of ill structuredness as a property arising from the requirement to resolve a large number of open constraints (i.e., parameters for which a value is not specified in the problem statement). Proceeding from this, we can say that each of the three aspects of a problem are ill-defined on the basis of the extent to which they are underspecified. Thus, the goal can be said to be ill-defined on the basis of the degree of uncertainty regarding what constitutes a solution to the problem. The materials can be said to be ill-defined on the basis of the uncertainty regarding what materials are relevant and/or available for use as part of the solution. The solution method can be said to be ill-defined on the basis of the degrees of freedom associated with the number of potentially relevant (legal) ways of manipulating the givens in order to attempt to formulate a solution. All of these elements can each contribute towards the ill-structuredness of the problem in general.

It is important to note that these dimensions each represent a continuum - there is no clear line between ill defined and well defined. Thus, it follows that the degree to which the overall problem is ill structured is also best represented on a continuum, meaning that, similar to the concept of design, most problems cannot be definitively characterized as well structured or ill structured, but merely as more or less structured. Another important point is that in making an assessment of how structured a problem is, the resources of the solver must also be considered (Simon, 1973; Voss and Post, 1988). To take an extreme example, if a solver is presented with the same problem twice under similar conditions, then while the problem may have been ill structured with respect to that solver the first time it was encountered, it is likely to appear much less so the second time (i.e., a large number of the open constraints will have been effectively closed when it was solved the first time).

2. Design problems have a large (often unenumerable) set of possible solutions (Logan and Smithers, 1993; Carroll et al., 1979; Rittel and Webber, 1973; Burns and Vicente, 1994).

3. Design problems are complex, consisting of many variables which can interact in multiple ways that are not wholly predictable (Goel and Pirolli, 1992; Carroll et al., 1979; Schon, 1988; Vicente et al., 1993).

4. Design problems do not have right or wrong answers, only better or worse ones (Rittel and Webber, 1973; Logan and Smithers, 1993; Goel and Pirolli, 1992). This often results in the fact that no clear stopping rule can be defined for a design problem - designers must

decide when a solution is "good enough". As Simon (1969) has pointed out, design problems are typically "satisficing" problems, not optimization problems.

5. Design problems have multiple, potentially conflicting requirements (Burns and Vicente, 1994; Logan and Smithers, 1993; Rittel and Webber, 1973). The quality of solutions can be judged from a variety of perspectives which designers must weigh against one another when developing a solution.

This is by no means an exhaustive list but represents some of the most commonly mentioned features of design problems. Many descriptions also attempt to define design problems in terms of the processes used to solve them, but the origins of most of these can be traced to the problem characteristics listed above. Some other features of design problems which are frequently noted can be attributed to the fact that these problems usually occur on a relatively large scale, involving many people, and taking anywhere from days to years to complete (Vicente et al., 1993; Goel and Pirolli, 1992). From this it can be argued that design is an inherently social process, with information distributed across people and organizations, where decisions tend to "emerge" from a process of social deliberation (Vicente et al., 1993). However, this does not negate the fact that, at some level, the act of individuals making decisions (or at least generating ideas) about how to arrange a group of entities in some sort of physical or possibly abstract informational space constitutes what would seem to be a meaningful unit of design activity. Although the process of design as an organizational activity may be regarded as an emergent property of large numbers of such decisions, how those individual decisions are arrived at, i.e., how those arrangements are devised, is surely important.

The view of design as a distributed, social process stresses the fact that many forms of design activity take place in open systems, subject to disturbances in the form of changing goals, requirements, and conditions. However, there are some tasks which have been characterized as design tasks which do not seem to share this property. These include music composition, writing, and painting (Goel and Pirolli, 1992).

As Goel and Pirolli (1992) have noted, design is too complex an activity to be adequately described in terms of necessary and sufficient conditions. Therefore, it seems fair to say that problems cannot definitively be described as "design" problems or "non-design" problems, merely as more or less "design-like", to the extent that they share commonly accepted characteristics of what defines a design problem, as described above.

As a footnote to the above discussion, it is worth mentioning that while some authors (e.g., Logan and Smithers, 1993) have asserted that design problems are

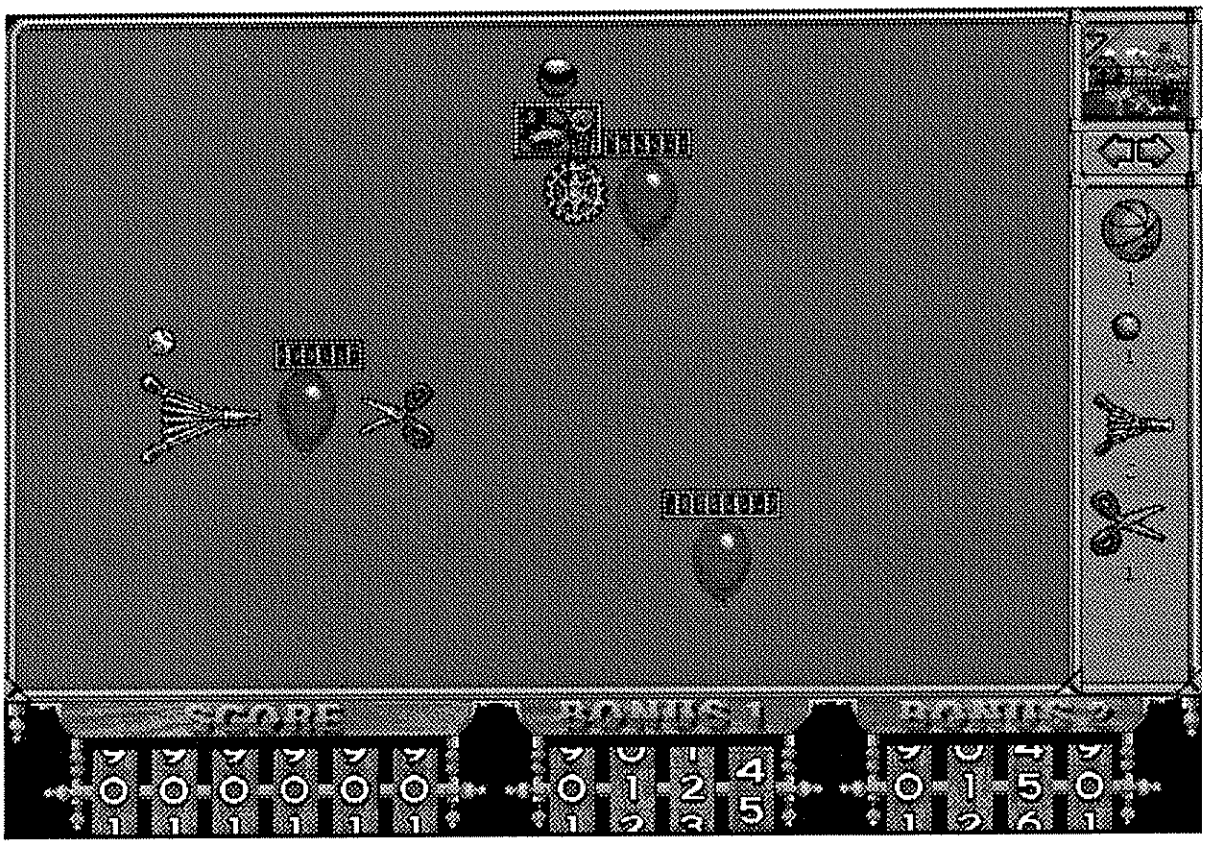
characterized by the fact that they are specifically not open to solution through purely search-based methods, Simon (1973) has argued that a significant proportion of behaviour observed in the study of complex, ill-structured problems, such as design problems, can in fact be described in terms of mechanisms identified in research on simpler problems (e.g., means-end analysis). Simon argues that the representation process for ill-structured problems may often consist of a well-orchestrated process of decomposition into several sufficiently independent and relatively well structured problems, each of which can then be attacked by the methods identified in research on simpler problems. However, while Simon describes an architecture for a problem solving system which can perform this decomposition, he has little to say about the methods by which the decomposition is performed, other than that they are likely to be a product of the problem solver's experience or background.

While the extent of the accuracy of Simon's viewpoint remains to be seen, it does not seem unreasonable to assume that continued research in relatively bounded domains will indeed contribute towards the understanding of problem solving behaviour in less well-defined domains such as design, thus bridging some of the gaps which seem to exist in our current understanding. To this end, there seems to be a need for research involving problem domains which are more complex than those used in some of the classic, scientifically oriented investigations, but which at the same time stop short of the immense complexity and ill-structuredness found in many humanistic studies of problem solving and design, thus making them more amenable to the formation of engineering oriented conclusions. This is where the current study attempts to fit in.

3.3 EXPERIMENTAL TASK CONTEXT

"The (even more) Incredible Machine" (TEMIM) is a commercial software product designed by Dynamix Incorporated. The premise behind TEMIM is straightforward. Users are given a series of scenarios, each of which consists of a goal, a set of pre-configured objects, and a set of user-configurable objects. The users' task is to configure the objects provided to them in such a way that the overall configuration satisfies the goal. An example will serve to illustrate the basic concept. Figure 1 represents a simple scenario as it would be initially presented to a user.

FIGURE 1. A simple TEMIM scenario.



The goal in this problem is given as "You must pop all the balloons." The pre-configured objects appearing in the largest portion of the screen (the "workspace") are fixed, and may not be moved or otherwise manipulated by the user. The objects appearing in the tall vertical window at the right of the screen (the "parts bin") are the objects which the user may, by clicking and dragging with the mouse, add to the workspace. The numbers below these objects represent how many of each are available. Therefore, in this case the user has one basketball, one tennis ball, two bellows, and one pair of scissors which they may add to the workspace and attempt to configure in such a way that all three balloons are popped.

The small window at the upper right corner of the screen (the "start machine" button) is central to TEMIM. Initially, when users are configuring objects within the workspace, time is "frozen", in the sense that all objects remain fixed in their positions in the workspace. When users click on the start machine button, time is "turned on" and the objects in the workspace become subject to the "laws of physics" with which TEMIM is programmed. It is through this process, hereafter referred to as "simulation", that users can test whether or not their current configuration satisfies the goal. If the current configuration satisfies the goal, the simulation stops automatically and the user is allowed

to proceed to the next problem. If the goal is not satisfied, the user must stop the simulation (by clicking the mouse button) and may then begin to manipulate objects in the workspace again. The simulation feature thus creates an iterative, discrete feedback loop providing information about various aspects of the current problem. Users may of course also mentally simulate the results of a (real or imagined) configuration. TEMIM allows users to use the simulation feature an unlimited number of times within each scenario.

3.3.1 Constraints in TEMIM

One way of describing TEMIM is in terms of the constraints which it forces users to operate under when attempting to solve a given puzzle. There are several sources of constraint in a given TEMIM problem. These include: the goal, the characteristics of the TEMIM "environment", the inventory of available objects, the affordances (i.e., potential uses; Gibson, 1979) of the available objects, the placement and initial state of objects within the workspace, and the relations between objects in the workspace. Each of these will be discussed below.

3.3.1.1 The goal. The goal is always clearly stated at the beginning of each scenario and is fixed for that scenario. The goals are always attainable, but the solutions are typically highly underspecified. There are often multiple qualitatively different potential ways of achieving the final goal involving various arrangements of various combinations of the available objects.

3.3.1.2 The TEMIM "environment". The environment of a given problem can be described in terms of the physics of the TEMIM world (which are fixed across problems) and the set of pre-configured objects in that particular problem.

The laws of physics, as modeled in TEMIM, represent a highly simplified, very rough approximation of reality, with some significant exceptions. Gravity is the primary force active in TEMIM, but not all objects are affected equally. Some objects can be placed anywhere in the workspace and will remain fixed and immovable during simulation episodes. Others will accelerate downwards in accordance with gravity until they are either stopped by another object or until they fall off the bottom of the workspace (the motion of objects is not constrained by the boundaries of the workspace). Objects also differ with respect to their solidity, meaning some objects allow others to pass through them, while some objects act as barriers to other objects. Most objects are subject to friction in some form (either sliding or rolling), with different objects possessing differing friction coefficients.

Collisions between solid objects can occur in two forms: a movable object colliding with another movable object, or a movable object colliding with an immovable object. In

collisions between movable objects, conservation of momentum seems to be approximately observed. When a moving object collides with an immovable object, the result depends mostly on the elastic properties of the moving object. Movable objects differ in terms of the degree to which kinetic energy is conserved in a collision (i.e., in terms of their "bounciness").

The second aspect to the TEMIM environment involves the pre-configured objects appearing in the workspace in each problem. These objects are fixed, both in terms of their position and initial state and cannot be altered in either respect by the user. During simulation however, they operate exactly as user-configured objects would. To solve problems, the user must usually integrate several of the user-configurable objects with the pre-configured objects in such a way that the resulting system of objects satisfies the goal. Thus, the set of pre-configured objects are an important part of most problems in that users must not only work around them but work with them to make them part of their solution.

TEMIM represents a closed, deterministic environment. If two simulations of the same configuration of objects are run, the results will be precisely the same. While it may be claimed that many of the properties described here are arbitrary to some extent, it is worth noting that they are far less arbitrary than the properties sometimes found in other computer environments. Moreover, a large part of solving problems in TEMIM consists of the discovery and exploitation of the constraints which these properties represent in the context of each new scenario.

3.3.1.3 Object inventories. A third source of constraint is the inventory of user-configurable objects. This inventory is fixed in each scenario, with limited numbers of a small set of object types available. However, this constraint is underspecified in the sense that not all of the objects given must necessarily be used to solve the puzzle. Often, "distractor" objects are included in the inventory, leaving the user to decide which objects are or are not useful within the scenario. In a problem with several potential solutions, some of the solutions may require all of the available parts to be used, and some may require only a few.

3.3.1.4 Object affordances. Another source of constraint can be found in the affordances of each of the objects. The affordances of a given object are limited and fixed across scenarios. However, these affordances may vary in terms of saliency and relevance, depending on the particular scenario. For example, a problem may require that a fishbowl be protected from a gun which is initially configured to fire at it. It turns out that a balloon can be used for this purpose if placed in the path of the bullet. While saliency of affordances cannot be objectively measured, it seems apparent that this would be an example of a non-salient relevant affordance for most people (the nominal, or "standard"

affordances of each TEMIM object are discussed below). Problems can prove very difficult to some users when a scenario calls for the use of an object in a new or unusual way (i.e., when less salient affordances become relevant to the problem). Moreover, the relevance of those affordances can change from scenario to scenario.

3.3.1.5 Object placement and initial state. Another source of constraint is that associated with the placement of objects. Objects may be placed anywhere in the two-dimensional workspace which is not occupied by another object. Some objects can be moved around in very small increments while others move in relatively coarse, discrete "jumps". In addition, many objects can be altered in terms of their initial state. Most objects can be "flipped" such that they face left instead of right, or vice versa (this changes the direction of rotation for rotating objects). Some objects, notably conveyor belts, ropes, belts, and barriers (see below), can also be stretched by a certain amount, making them longer or shorter within certain fixed limits. Exactly where objects can be placed in a given scenario and how they can be altered is left up to the user to discover through trial and error.

3.3.1.6 Relations between objects. A final source of constraint can be found in the form of the relations between objects. In a given scenario, these relations emerge in the form of cause-effect chains as a result of the placement, initial state, and characteristics of the objects concerned. Relations can also be found at higher levels where objects are combined into systems of objects which exhibit novel affordances and relations with other objects or systems of objects. Thus, both affordances and relations emerge in a combinatorial fashion out of the particular configuration of objects in a given scenario. The essence of solving problems in TEMIM lies in discovering and manipulating the relations between objects and systems of objects by assembling them into functions, so that the resulting system meets the prescribed goal.

3.3.2 TEMIM Objects

The inventory of objects in TEMIM is quite extensive, consisting of 55 different types of objects each with their own unique characteristics. Of these 55, only 43 were used in the problem presented to subjects in the current study. It is important to emphasize that not all objects are used in each scenario. Rather, individual scenarios each use a small subset of object types, with varying quantities of each of these object types available. The 43 objects used in this study can be roughly classified based on their standard functions as described in the user's manual (Dynamix Inc., 1993). The general classes are:

1. "Kinetic" objects. These objects act basically as carriers of kinetic energy. This class includes: 5 types of balls (baseballs, basketballs, bowling balls, cannonballs, and tennis balls), each with different inertial and elastic properties, as well as balloons, buckets and cages.

2. Objects which transform energy from one type to another. These objects can also each be viewed as a unified tandem of an energy sink and an energy source, but seem to merit a unique classification. This class includes: generators (rotational to electrical), electric motors (electrical to rotational), solar panels (light to electrical), fans (electrical to wind), conveyor belts (rotational to kinetic), jack in the boxes (rotational to kinetic), and windmills (wind to rotational).

3. Objects which transfer forces. These objects can act as intermediaries in the transfer of forces from one object to another. They can alter forces in terms of direction and/or location. This class includes: ropes, pulleys, belts, gears, and teeter-totters.

4. Objects which are sources of energy. These objects provide energy in a specific form when actuated properly. This class includes: light switch/outlets (electrical, actuated by impact from a kinetic object which flips the switch), flashlights (light, actuated by impact on switch from kinetic object), lightbulbs (light, actuated by tension force transferred through a rope), boxing gloves (kinetic, actuated by impact against button from a kinetic object), guns (kinetic, actuated by tension force on trigger transferred through a rope), trampolines (kinetic, adds energy to objects which fall on it), cannon (kinetic, actuated by focused light or flame), dynamite (explosive, actuated by focused light or flame), rockets (kinetic, actuated by focused light or flame), candles (flame, actuated by focused light or flame), dynamite with plunger (explosive, actuated by impact on plunger by kinetic object or by tension force transferred through a rope), mouse cages (rotational, actuated by impact from kinetic object), monkeys on exercise bicycles (rotational, actuated by downward tension on blind transferred through a rope).

5. Objects which act as barriers. These objects act primarily to absorb and/or redirect the kinetic energy of other objects. These include: brick walls, wooden walls, pipes, and inclines. They differ primarily in terms of how they are affected by explosions.

6. Other. These are the remaining objects which do not fit into any of the categories described above: scissors (cutting action, actuated by impact from above or below by a

kinetic object), magnifying glasses (focuses light energy), cats (kinetic/"intentional" objects, attracted to mice and broken fishbowls), mice (kinetic/"intentional", repelled by cats) and fishbowls (attract cats when broken, actuated by impact from a kinetic object)

To reemphasize, these descriptions are based merely on the standard affordances for each object as might be deduced from the descriptions listed in the user's manual. Many of these objects also have one or more less obvious affordances (e.g., the exhaust from a rocket can be used to burst a balloon). Therefore, many of the objects may actually belong in several of the categories described above. In fact, as alluded to earlier, it is often advantageous and occasionally necessary to use one or more objects in non-obvious ways in order to solve certain problems.

To summarize, problems in TEMIM, like the Apollo 13 scenario described in the prelude, require subjects to accomplish a specific goal by integrating a number of objects from a small set with an arrangement of pre-configured objects. While subject to various sets of constraints defined by the TEMIM environment and the form of the specific scenario, there are a practically infinite number of legal ways of configuring the given objects, although only a small subset of these will constitute solutions. Perhaps the key feature of problems in TEMIM is that they are inherently concerned with functions of objects and systems of objects, which are <u>not</u> uniquely defined and can be viewed as emergent properties dependent on the precise assortment and configuration of objects. For example, given a flashlight and a magnifying glass, the affordance of lighting a candle is not a property of either object individually, but can only be associated with the system of objects that the flashlight and magnifying glass together represent. In fact, the relevant system in this example could be extended to include whatever object serves to hit the button on the flashlight to switch it on. Moreover, as discussed previously, the saliency and relevance of affordances of objects or systems of objects can change depending on the scenario and/or the solution method being attempted.

3.3.3 Do Problems in TEMIM represent Design Tasks?

Given the above list of typical features of design problems, an attempt can now be made to assess to what extent TEMIM can be characterized as a design task. Each point will be discussed individually.

<u>1</u>. To assess the degree to which TEMIM is ill-structured, each aspect of the problem, that is the goal, the givens, and the solution methods need to be considered separately. It is clear that the goal in any given problem is <u>not</u> ill defined. The goal is always stated

explicitly at the beginning of each problem and there is a binary evaluation mechanism (the simulation feature) which indicates whether or not a solution has been found. The givens in each problem are well defined in the sense that there is no ambiguity concerning what materials are available for use in a solution attempt, but they are ill defined in the sense that there is very often a large amount of uncertainty over which of the available objects are relevant for solution of the problem. This is compounded by the fact that many of the objects may have multiple potentially relevant affordances. Finally, the nature of how the givens may be utilized to attempt a solution is quite ill defined. While the constraints on placement and initial state must be respected, there are clearly a large number of degrees of freedom concerning the systems which may be created out of the available materials. Overall then, it can be said that problems in TEMIM are partially ill structured (especially when compared with tasks such as cryptarithmetic, symbolic logic, tower of Hanoi problems, and others used in previous laboratory studies of problem solving).

2. With respect to the number of possible solutions to a given problem, many scenarios in TEMIM allow for several potential configurations which will achieve the goal. It is not usually possible to say definitively how many different arrangements of the given objects will accomplish the goal, although there is clearly often more than one. Thus, TEMIM seems to conform to this feature of design problems as well.

3. Problems in TEMIM, while obviously not as complex as many problems typically encountered in design, do consist of a relatively large (but bounded) set of variables which can interact in unexpected ways. While the environment is deterministic and not subject to external disturbances, it is complex enough that it would be very difficult for a user to accurately predict all aspects of the outcome of any given configuration of objects. This is the reason that the simulation feature is so important as a source of information in TEMIM. Thus, TEMIM seems to exhibit this feature of design problems as well, albeit at a relatively small scale.

4. Due to the well-defined nature of the goals and the fact that an unambiguous test of the adequacy of a solution exists in the form of the simulation feature, solutions in TEMIM can definitively be labeled as right or wrong. Therefore, in contrast with design problems, TEMIM scenarios do not contain any uncertainty surrounding the sufficiency of solutions. However, given the fact that there are often multiple ways of solving the problems, it may also be possible to distinguish solutions as "better" or "worse". For instance, it would be possible to judge solutions on the basis of how long subjects take to produce a solution,

how many objects were used, how many times the simulation feature was used, or how many different solutions were attempted. These all have natural analogs in other design problems, for instance: time used on a project, the amount of material used, the amount of effort spent generating or searching for information, and the number of designs generated, respectively. In all cases, less would be considered better. Although subjects were not explicitly given any of these criteria in the current study, they were told that all of these factors, with the exception of the last one, would be measured. In any case, it seems reasonable to expect subjects to adopt some of these criteria spontaneously.

5. The only explicit requirement users must be concerned with is to solve the problem, with the only criteria being the judgment of the simulation feature. TEMIM therefore does not explicitly exhibit the large number of potentially conflicting requirements typical of design problems. However, proceeding from the discussion of the previous point, problems in TEMIM can be said to have multiple, conflicting requirements to the extent that the criteria subjects adopt conflict with one another.

From the above discussions, if nothing else, it seems clear that it is very difficult to definitively state how much a given problem qualifies as "design-like". With respect to the common features of design problems, TEMIM at least partially exhibits most of these features to some degree, subject to the limitations imposed by its scale and by the fact that it represents a closed system with very definite boundaries. So while there can be some debate about how design-like TEMIM actually is, it would be difficult to make the argument that TEMIM is not design-like.

As an additional point, it bears noting that, in terms of the type of medium in which design takes place, TEMIM shares one general characteristic with many design tasks in that it represents a virtual world. Schon (1988) and Goel and Pirolli (1992) have noted that designing consists not of the building of artifacts, but of the construction of representations of things to be built. Things such as sketches, diagrams, and models serve as virtual worlds in which designers can explore their ideas at low cost and with relatively little risk. In this sense at least, TEMIM clearly embodies a typical feature of many design tasks.

To summarize then, TEMIM represents an environment which offers the opportunity to study design problem solving in a relatively controlled setting[1]. While

[1]Rasmussen (1987) has argued that computer games may in fact be a highly appropriate setting in which to study cognitive behaviour. Speaking from the perspective of user modeling for applied settings, Rasmussen argues that computer games afford the opportunity for subjects to become highly skilled in a relatively short period of time, allowing for the observation of a full range of cognitive behaviour, while still allowing for experience to be relatively closely controlled.

clearly representing only a relatively narrow class of design problems, and while not fully exhibiting all of the characteristics most typically associated with design problems, it is felt that it retains enough of the character of design to serve as a type of "design microworld".

4. PILOT STUDY

Because of the exploratory nature of this study, an informal pilot study was undertaken first to attempt to narrow down the phenomena of interest and identify factors potentially relevant to performance. A brief summary of the pilot study and the results is presented below.

4.0.1 Subjects

Initially, one of the potential factors of interest identified was educational background. Thus, the subjects were deliberately chosen to be heterogeneous in this respect. The backgrounds of the four subjects participating ranged from applied arts (interior design) to applied science (electrical engineering, computer science). One subject had training in both fine arts and applied science. No attempt was made to control for sex differences; two of the subjects were male, two were female. Subjects were paid $40 each for their participation. A summary of the subjects used in the pilot study is presented in Table 1 below.

Table 1. Pilot Study Subjects

SUBJECT	BACKGROUND	SEX (M/F)
BH	Computer Science	F
LD	Electrical Engineering	M
MJ	Systems Engineering / Fine Arts	M
SF	Interior Design	F

4.0.2 Apparatus

"The (even more) Incredible Machine" (version 1.0) was run on a Apple Quadra 660AV computer with a 15-inch colour monitor and standard Apple Desktop Bus Mouse. TEMIM was run with the sound and music options switched off to avoid distracting the subjects. Verbal protocol collection was performed using a Sony CCD-TR81 Hi-8 Handycam with an external clip-on microphone.

4.0.3 Procedure

The experiment consisted of two sessions conducted on consecutive days. The first session included presentation of all of the introductory materials. Subjects were asked to

read an overview of the experiment and then to fill out an informed consent form and demographic questionnaire. They were then given instructions about how to operate TEMIM and a brief set of instructions about giving verbal protocols (copies of the introductory materials are included in Appendix A). Following this, subjects performed all 21 of the "tutorial" problems included in the TEMIM software package. These are problems which are intended to introduce users to the TEMIM environment and the various objects used in the scenarios. A time limit of 15 minutes was imposed on all of the tutorial problems. If subjects failed to solve a problem in this time, they were then shown a solution (the same solution was always shown for specific problems). The second session consisted of 11 further trials, which were intended to be more challenging. These trials had a time limit of 30 minutes and no solutions were given for uncompleted trials. Subjects were asked to provide verbal protocols as well as post-hoc explanations of their solutions for all trials. In addition to recording verbal protocols, a videotaped image of the computer screen was recorded for all trials.

One of the questions that the pilot study sought to answer was methodological. It had been hypothesized that subjects' verbalizations might be more revealing if, rather than operating TEMIM themselves, they were forced to direct the experimenter in operating the game. It was thought that this might promote more statements regarding the reasons for subjects' choice of actions. Therefore, for every second trial in the second session, the subjects were instructed to operate TEMIM themselves while for the other half of the trials, the experimenter operated TEMIM and the subjects were asked to direct the experimenter to perform part placement, simulation, etc.

In order to try to separate effects due to whether the subject or experimenter was operating TEMIM from the influence of differences between problems, three repeated problems were used. The occurrences of the repeated problems were always eight trials apart. For one repeated pair, the subjects first performed the trial by themselves, but were asked to direct the experimenter for the second occurrence. For another pair, subjects performed both trials while controlling TEMIM themselves. For the final pair, subjects performed both trials by directing the experimenter. Verbal protocols were collected for all trials.

4.0.4 Performance Measures

Data were collected for all trials (from both sessions), including trial completion times, when parts were placed in or deleted from the workspace, and when the simulation feature was used. Aside from quantitative performance differences (i.e., trial completion times), the study also sought to identify strategy differences, particularly those associated

with subjects' use of the simulation feature. Patterns in subjects' use of the simulation feature were examined carefully.

4.0.5 Summary of Results

Because of the small sample size, many of the results could only be suggestive rather than conclusive. However, attempts were made to confirm as many of the results as possible using appropriate statistical analyses.

The methodological issue outlined above was examined first. Informally, it was noted that during trials on which the experimenter operated TEMIM, subjects seemed in fact to make less verbalizations regarding their own mental states and motivations, and to spend more time instead giving detailed instructions to the experimenter about exactly where to place parts, how to orient parts, etc. The results suggested that during the second session, subjects' tendencies to use the simulation feature may have been inhibited on experimenter-controlled trials. This matched well with the informal observations made during the sessions. Therefore, it was decided to discard the experimenter-controlled trials from subsequent analyses.

The validity of the data from the repeated problems was also called into question by the informal observations made during the data collection session. Subjects typically displayed very good memory for solutions. Therefore, when encountering a problem for the second time, the time and number of simulations required to solve the problem usually dropped drastically. However, if the subject had not solved the problem during its first appearance, despite remembering the problem well, their difficulty typically continued during the second attempt. Due to the small sample size, the analysis did not confirm these suspicions conclusively, but the results were suggestive enough that it was decided that the data from the second occurrence of each of the repeated problems should also be excluded from further analyses.

A Friedman two-way analysis of variance by ranks (Siegel and Castellan, 1988) was performed using data from all of the trials not discarded due to the reasons discussed above. The non-parametric analysis was chosen due to the fact that the trials from the first session were performed under a 15 minute time limit while the trials from the second session were performed under a 30 minute time limit. The assumption of a normality in the distributions was therefore obviously not warranted. The dependent variables examined were trial completion time, number of simulations, rate of simulations, and the number of parts used in each trial. Pairwise comparisons were performed with α-values scaled according to the number of comparisons as recommended by Siegel and Castellan (1988).

The results showed significant subject effects for all of the dependent measures except for the number of parts used. Differences between trials were not examined, as no methods for systematically classifying relevant features of problems were available. Effects due to problems were therefore not pursued.

Pairwise comparisons between subjects' ranks in completion times revealed that subject LD was slower than any of the other subjects. The remaining subjects did not differ significantly from one another. However, examination of each subject's average completion times suggested that MJ and SF were faster than BH, who was in turn faster than LD. These times are shown in Table 2 below.

TABLE 2. Average completion times by subject.

SUBJECT	AVG. TIME (sec)
BH	411
LD	489
MJ	324
SF	319

Pairwise comparisons of subjects' ranks with respect to the number of simulations was expected to be uninformative because these ranks were likely to be confounded with completion time. This analysis revealed that MJ used significantly less simulations than any of the other subjects. None of the other subjects were significantly different from one another.

The more revealing measure with respect to subjects usage of the simulation feature was the rate of simulation. The results of the pairwise comparisons revealed two distinct groups within the data. Subjects MJ and LD were not significantly different from one

TABLE 3. Average rate of simulations by subject.

SUBJECT	SIMULATIONS (per 100 seconds)
BH	3.57
LD	1.84
MJ	1.84
SF	3.44

another, but both differed from BH and SF, who also did not differ from one another. Examination of the numerical values reveals this pattern clearly (see Table 3).

These results were curious because, counter to expectations, they seemed initially to be unrelated to the results from the analysis of completion times. For example, MJ and LD, although nearly identical with respect to simulation frequency, had widely different average completion times. Clearly, the differences in completion time could not be directly attributed to subjects' tendencies to gather feedback by use of the simulation feature. In addition, these results suggested that a possible difference between male and female subjects.

It was suspected that subjects might differ with respect to not only how much they used the simulation feature, but also with respect to patterns of usage within trials. In order to test this, a Kolmogorov-Smirnov test was performed. All trials were normalized with respect to completion time then partitioned into 40 equal segments. The total number of simulations across all trials in each segment was then calculated for each subject. The results revealed no significant differences between subjects in the distribution of simulations during trials. The performance differences could therefore also not be attributed to consistent differences in simulation strategies.

A number of similar qualitative analyses were performed in order to attempt to identify patterns in the use of the simulation feature that might account for the performance differences observed. None, however, revealed any consistent patterns. One of the alternative explanations proposed was that the differences in performance might be mediated by spatial ability. Because the problems are inherently spatial in nature, involving the arrangement and movement of objects within a two-dimensional space, low spatial ability might significantly hamper a subject's performance in two ways. Firstly, poor ability to visualize the layout of parts could lead to problems in composing the spatial arrangement of parts in a given problem. Secondly, the ability to visualize the dynamics of an existing arrangement (i.e., mental simulation ability), would presumably also be correlated with spatial ability.

4.0.5.1 Spatial ability test. In order to test the hypothesis that spatial ability might be a mediating factor in performance, the four subjects were called back after the experiment and asked to complete the VZ-2 test for spatial visualization (Ekstrom, French, Harman, Dermen, 1976). This test had been found to be highly predictive of performance differences in a previous pair of studies involving a spatial task (Vicente, Hayes, and Williges, 1987; Vicente and Williges, 1988). The test consists of two similar parts, each with ten multiple choice questions. Each question depicts a square piece of paper which, in a series of 2 to 4 drawings, is illustrated being folded in a certain way. A single hole is

then illustrated to be punched in the folded piece of paper. Subjects are asked to imagine this procedure and to then decide from among five choices what the piece of paper will look like (i.e., where the holes will be) when it is unfolded. Subjects are given three minutes for each ten-question section. The test was scored by giving one point for each correctly answered question and by subtracting 0.2 for each incorrectly answered question. Subjects were informed of the scoring scheme in advance. The results are presented in Table 4.

TABLE 4. Subject scores (out of 20) on spatial ability test.

SUBJECT	SCORE
BH	12.6
LD	8.6
MJ	19
SF	15.8

These results seemed to shed a great deal of light on the differences previously described. The drastic difference in spatial ability between MJ and LD may be the reason that although they simulated roughly equally frequently, MJ outperformed LD by a large margin. Similarly, SF's slightly higher spatial ability score may partly explain why she was, on average, over one and a half minutes faster than BH, although exhibiting a similar simulation frequency. Finally, the fact that SF simulated much more frequently than MJ might explain why, although her spatial ability score was lower than that of MJ, she was able to achieve a very similar average completion time.

It is interesting to note that the two subjects with arts backgrounds were of high spatial ability while the two subjects of lower spatial ability had exclusively applied science backgrounds. This suggested that it might in fact be more directly relevant for the subsequent study to classify subjects based on spatial ability rather than educational background.

To attempt to examine subjects' verbalizations in detail, three trials were chosen for transcription on the basis of wide variability in performance between subjects. It was hoped that the differences might be partially accounted for by features in the verbal protocols. Each of these three trials was fully transcribed for each subject. Subjects' statements were classified among four levels of abstraction, and the number of goal statements was counted. The results of these analyses proved inconclusive; no consistent differences between subjects were observed with respect to verbalizations at the various abstraction levels or with respect to the number of goal statements.

Most of the subjects' verbalizations during the trials were classified at the two lowest levels of abstraction (physical form and physical function; Rasmussen, 1986). Very rarely were statements made at higher levels of abstraction. However, it was noted that during the post-trial solution explanations, which were not transcribed, subject SF and, to a lesser degree, subject MJ tended to make relatively more statements than subjects BH or LD at the abstract function level (Rasmussen, 1986). While strictly an informal observation, this suggested that SF and MJ may have been more prone to think in these terms, which may have served as an additional factor in their superior performance with respect to completion times.

The final analysis to be reported from the pilot study concerned the relationship between subjects' frequency of use of the simulation feature and the complexity of the problems encountered. As mentioned previously, no psychologically relevant way of classifying or ranking the problems with respect to complexity or difficulty was known. However, for the purposes of these analyses, the number of parts available was used as a surrogate. This refers to the number of parts included in the parts bin (i.e., available for subjects to add to the workspace). This was found to be negatively correlated with completion time for most subjects, and therefore seemed a suitable method of estimating problem complexity. The results revealed that for subjects MJ and LD, frequency of simulation seemed to remain relatively constant regardless of the number of parts. However, for subjects BH and SF, the rates of simulation seemed to be negatively related to the number of parts provided. For problems with low numbers of parts their rates were much higher than those of MJ and LD, but as the number of parts increased, their rates became smaller and smaller. For the problems with the highest numbers of parts, their rates approached those of MJ and LD. These results suggested that the strategies of BH and SF, at least with respect to feedback generation, may have somehow been sensitive to the complexity of problems, whereas those of MJ and LD were not.

4.0.6 Conclusions

The main contributions of the pilot study were related to the methodological issues explored and to the interaction between spatial ability and frequency of simulation. With respect to the former, it was concluded that it was best to allow subjects to operate TEMIM themselves. The method of using the experimenter as an intermediary proved to be too awkward and distracting for subjects and seemed to detract from subjects' verbalization of their thought processes. Also, the influence of spatial ability was found to be potentially highly relevant to performance, more directly perhaps than educational background. With respect to simulation frequency it seemed that there were indeed distinct differences

between subjects in terms of their tendency to use the simulation feature as a source of feedback. The effect that this tendency had on performance seemed to be mediated by spatial ability. While no effects due to the level of abstraction of subjects' statements were found for the trials examined, informal observations suggested that this factor might nonetheless be relevant. Finally, simulation frequency seemed to be related to problem complexity, although this effect was only present for certain subjects.

5. EXPERIMENTAL FACTORS AND MEASURES

The pilot study helped to point towards a number of factors which seemed to be relevant to either predicting or measuring the performance of subjects solving problems in TEMIM. Clearly, spatial ability stood out as a potentially relevant performance shaping factor. Simulation frequency was identified as a potentially discriminating measure of subjects' interaction with problems. Although no consistent differences between subjects were identified with respect to verbalizations for the three trials examined in the pilot study, it was felt that, given some of the suggestive informal observations collected and the design-like nature of problems in TEMIM, the two categories of verbalizations measured (abstract references and goal statements) remained as the most potentially fruitful measures of subjects' verbal activity relevant to performance.

In order to more clearly define these factors and their hypothesized relevance to performance in TEMIM, each will now briefly be discussed in more depth and an attempt made to situate them in the context of problems of the type encountered in TEMIM. In addition to spatial ability, constraint level is introduced as a second experimental factor and its motivation and relevance are discussed.

5.1 FACTORS

5.1.1 Spatial Ability

The results of the pilot study suggested that spatial ability might be a highly relevant factor shaping performance in TEMIM. Many authors (e.g., Hegarty, 1991; De Kleer and Brown, 1983) have identified relationships between spatial reasoning ability and performance on similar problems, with a particular emphasis on the role that spatial visualization can play in allowing subjects to manipulate spatial representations of systems in order to predict their behaviour.

For example, Hegarty (1991) asserts that in order to infer the behaviour of any mechanical device, subjects rely on several types of knowledge, in concert with spatial representations of the configurations of objects, to "mentally simulate" the dynamic and kinematic relations between components. Because mechanical problems fundamentally involve the motion of objects through space, Hegarty hypothesizes that subjects are likely to make extensive use of spatial representations of the problems. These representations serve as an index to information, since interactions between components depend on their relative locations in space. Secondly, and more importantly for the current study, they

allow subjects, in combination with relevant knowledge, to infer a qualitative causal model of the system and thus derive information about how the system will behave while in operation. The ability to construct and manipulate approximately correct causal models of systems (existing or envisioned) would certainly seem to be highly relevant to performance in TEMIM. In fact, Hegarty notes that this process is central to the design of mechanical systems, a point which has also been made by Simon (1969) in discussing the spatial nature of design.

5.1.2 Constraint and Knowledge of Constraint

As discussed in section 3.2, the degree to which a problem is ill-defined is normally thought of as inversely associated with the amount of constraint inherent in the initial state of the problem. Because a relative lack of constraint is so often identified as a characteristic feature of design problems (i.e., they are characteristically ill-defined), it was felt that it might be very informative to attempt to manipulate the amount of constraint in problems, and/or subjects' knowledge of problem constraints, as a factor in the study. Such manipulations, it was expected, would sharpen performance differences and perhaps lead subjects to adapt their strategies under reduced-constraint conditions in ways which could be captured by the measures outlined below, and which might thus suggest how designers cope in the face of relatively ill-defined problems. The precise forms of constraint manipulations performed in this study are discussed in section 6.5.

5.2 MEASURES

5.2.1 Simulation Frequency

Simulation frequency is a phenomenon relatively unique to the TEMIM context. It can however be viewed generally as an instance of subjects' tendency to request feedback about the status of their current designs. The pilot study revealed that this measure might be related to performance through the mediating influence of spatial ability. In particular, it appeared that spatial ability and simulation frequency might trade-off in their effects on performance. Thus, subjects of lower spatial ability who simulated relatively frequently might be able to achieve performance similar to subjects of higher spatial ability who simulated less frequently. It is possible then, that subjects of lower spatial ability could be expected to adapt their usage of the simulation feature to try to compensate for their presumably reduced ability to make inferences of the type described in the previous section.

5.2.2 Abstraction

Abstraction can be a fruitful method of coping with the complexity of difficult problems. Viewing a problem from a more abstract perspective is a means of simplifying the representation, effectively reducing the size of the problem space and allowing the solver to focus on essential aspects of the problem. The sense of abstraction used here is that exemplified by Rasmussen (1985). In Rasmussen's abstraction hierarchy (AH), a construct used to model complex work (problem solving) environments, levels of abstraction differ not just in the amount of detail revealed, but also in the very concepts used to describe the domain[2]. At the lowest levels, the elements of the environment are described in very concrete terms (e.g., their physical location and appearance). Moving up through levels, the system is described in terms which specify the functions of elements or subsystems of elements in increasingly abstract terms (i.e., the representation maintains less and less of the physical basis for the functions being described). The highest level in the AH describes the purpose or objectives of the system as a whole. Thus each level describes the system completely, but the language at each level will be entirely different.

It is possible to view practical problem solving as a search through an AH. Vicente and Rasmussen (1992) cite several examples of studies where protocols from problem solving episodes have been mapped onto AH representations of the domains. The important point is that because the levels are linked through means-end relations to the purposes of the system, moving between levels can be an efficient way to engage in goal-directed problem solving. In diagnosis tasks for example, by starting at the higher levels and following the means-end links from the affected functions down through the hierarchy, the focus of attention can be constrained and an effective "zooming-in" on the locus of the disturbance can be performed. Vicente, Christoffersen, and Pereklita (1995) found that, for subjects performing a diagnosis task in a process control system, performance was positively correlated with the use of such a strategy.

For design-like problems, the case is slightly different. In the example just cited, the AH describes an existing system, allowing subjects' reasoning to be represented as a path through that particular AH. In design, there is initially no complete system to be described. The problem is to construct a system, given a goal and a more or less well specified set of initial conditions and resources. To take the example of TEMIM, subjects are given the objective of the system to be constructed (i.e., the top level in the AH) and a

---

[2]This may be compared with the IP view of abstraction, exemplified in the work of Newell and Simon (1972), and Sacerdoti (1974). In this view, abstraction is defined as a method for disregarding certain details of a problem which are deemed temporarily inessential to the solution process. This can be characterized as a clever method of temporarily "skipping steps" in devising the causal structure of elements within a single level of an AH.

set of materials (i.e., the elements which will make up the lowest level in the AH). The process of solving problems in TEMIM can be viewed as the construction of an AH which connects the given elements at the lowest level to the goal at the highest level. Note that this is true regardless of the reasoning process actually used by subjects to arrive at a solution. The point to be made here is that, as a heuristic for solving even isolated subproblems, the type of abstraction embodied by the AH can be a powerful tool to guide reasoning in pursuit of goals. Therefore, even if subjects do not reason about problems as a whole in this manner, they may solve subproblems in a similar way, representing certain aspects of the problem in abstract terms to constrain the search for solutions.

5.2.3 Goal Chaining (Causal Reasoning)

It is important to differentiate between the functional reasoning described in the section above (i.e., functional abstraction) and causal reasoning, which occurs within a level of abstraction. Functional abstraction is a method for representing elements of a system in more or less abstract terms; the elements (i.e., functions) are recast in terms of different concepts. Causal reasoning, on the other hand relates to the causal structure of elements within a level of abstraction. This type of reasoning may still be concerned with functions, expressed in the language of the level of interest, but is concerned with how the functions within that level relate to one another in terms of causes and effects.

Two of the most basic methods which can be used for causal reasoning are forward chaining and backward chaining (see e.g., Newell and Simon, 1972). Both are basically associative methods where the action to take next is determined by the current state of the problem. However, there is an important difference in that for backward chaining, the influence of the current causal goal is explicit, while for forward chaining it is implicit (if it is present at all). In backward chaining the reasoning process works backwards through a causal sequence from a specific goal to the givens of the problem by looking for actions (operators) which will achieve the goal. If one is found, it is applied inversely, and a new goal of arriving at the resulting state is set. In working forward, the reasoning process works by examining the current state to evaluate which actions can possibly be taken as the next step in the causal sequence. No goal needs to be active, other than the obvious goal of looking for an operator to apply to the current state.

The first process, backward chaining, seems as though it would be a more effective way of solving problems in TEMIM, at least when viewed at a holistic level. Because the problems have very definite goals, building the causal structure of the solution (at any given level of abstraction) by working backwards would appear to be a much more effective way of constraining the search for relevant operators (individual or functional

groups of objects). Except for problems that are highly constrained, working forward would, in general, be an inefficient method of constructing systems. As subjects become more experienced, they may be expected to exhibit an increased tendency to engage in episodes of forward reasoning for familiar subproblems. However, in devising the overall structure of subsystems necessary to form a final solution, backward chaining would still likely be relevant. There are a number of additional factors which could also influence the appearance and usefulness of working forward or backward (e.g., use of analogies or abstract concepts to guide solutions, or use of feedback from the simulation feature).

With this small set of factors and measures proposed as potentially relevant to shaping or capturing aspects of performance in TEMIM, it is possible to pose some of the general questions which the current study attempts to explore. Firstly, how do the factors above relate to performance? How do they interact? In terms of these factors, what separates subjects who perform well from those who perform relatively poorly? How do subjects adapt in the face of constraint manipulations?

It is important at this point to reemphasize the exploratory nature of this study. The focus is intended to be broad rather than deep; the purpose is primarily to generate questions, not answers. Thus, while the current study seeks to start down the path towards answers for some of the general questions outlined here, the intent is not necessarily to arrive at definitive answers here. Indeed, the study would be considered a success largely to the extent that new and interesting forks in the path are discovered along the way.

6. METHOD

6.1 EXPERIMENTAL DESIGN

Subjects were divided into four groups based on the scheme illustrated in Figure 2. All subjects performed the same 16 trials, consisting of 4 blocks of 4 trial types, each corresponding to a specific type of constraint reducing manipulation (see below). Subjects in the reduced constraint group performed trials in their reduced constraint format, while the baseline group performed them in the standard format (to be explained below).

FIGURE 2. Subject Groups

		TREATMENT GROUP	
		BASELINE	REDUCED CONSTRAINT
SPATIAL ABILITY	LOW	4 subjects	4 subjects
	HIGH	4 subjects	4 subjects

6.2 SUBJECTS

Sixteen subjects were used, with four subjects of each level of spatial ability assigned to each of the two treatment conditions. The results of the pilot study revealed possible sex differences, so only male subjects were used in this study. The subjects ranged in age from 19 to 34 years of age and were selected on the basis of educational background and spatial ability. Potential subjects were rejected if they had used or seen TEMIM previously. Only subjects with science or engineering backgrounds were chosen so as to attempt to control, at a very coarse level, for possible biases in problem solving style resulting from educational background (effects of this type have been previously reported in a design-like task by Lawson, 1984). Subjects were screened for spatial ability and divided up into the two spatial ability groups using the VZ-2 test of spatial visualization

ability (Ekstrom, et al., 1976; described in section 4.0.5). A summary of the subject groups is presented in Table 5 below.

TABLE 5. Subjects (HS denotes High Spatial; LS denotes Low Spatial; 1 denotes baseline group; 2 denotes reduced constraint group).

SUBJECT	BACKGROUND	VZ-2 SCORE	AGE
Group HS1			
CS	Mechanical Engineering	18	24
JB	Mechanical and Computer Engineering	18	26
KC	Electrical Engineering	18.8	23
SJ	Computational Neuroscience	20	25
Group HS2			
FN	Computer Science and Engineering Science	18.8	19
JH	Mechanical and Biomedical Engineering	18	25
MT	Aerospace Engineering	18.8	22
TB	Aerospace Engineering	20	21
Group LS1			
DG	Computer Science	12	21
JG	Electrical Engineering	13.4	24
ML	Physiology and Zoology	13	25
PK	Physiology and Geology	13.8	26
Group LS2			
AS	Computer Science	10.6	23
DN	Mechanical Engineering	8.4	21
JO	Electrical Engineering	13.6	34
PP	Computer Science	12.8	26

6.3 APPARATUS

All apparatus was identical to that used for the pilot study. Please refer to section 4.0.2 for details.

6.4 EXPERIMENTAL TASK

The only objective given to subjects was to complete the problems. They were informed that completion times, part usage, and "various other measures" would be collected, but were not given any specific performance targets. Subjects were not given any information in advance about the specific types of problems they would encounter.

6.5 TRIAL TYPES

Trials were performed under four distinct sets of conditions, each of which will be described below. The differences between the treatment groups lay in the conditions under which they were asked to perform trials. With the exception of "normal" trials, the reduced constraint group (hereafter referred to as Group 2; the baseline group will hereafter be referred to as Group 1), performed all trials under conditions where the structure (i.e., constraint) inherent in the problem was reduced or hidden in some fashion. All trials consisted of problems supplied with the TEMIM software package, although some were slightly modified (see below). Problems were chosen based on the results of the pilot study and on perceived difficulty so as to be neither trivial nor too difficult for the majority of subjects. For all data collection trials, a 30 minute time limit was enforced. Snapshots of the initial configurations of each trial used during the data collection sessions are included in Appendix B.

6.5.1 Normal (N) Trials

Under this condition, problems were presented in their original unaltered form and there were no restrictions on either group aside from the 30 minute time limit.

6.5.2 Covered Bin (CB) Trials

Under this condition, the level of constraint was reduced by not allowing subjects the opportunity to observe a priori what parts were available for completing the problem. For Group 2 subjects, an opaque piece of paper was taped onto the right side of the screen, covering the parts bin. When a subject wanted to add a part to the workspace, he had to ask the experimenter for the specific part that he wanted by name. Because the number of possible types of parts available is quite high (43), the subjects were given an alphabetically ordered list of all of the parts that could possibly be available as a memory aid. If any

instances of the part the subject requested were available, the experimenter would ask the subject to turn his head while he (the experimenter) retrieved it from the parts bin and placed it in an open part of the workspace. The subject was then allowed to place the part as he wished within the workspace. If no instances of the part the subject requested were available, the experimenter would inform the subject of this. Subjects were not allowed to mark the list of parts in order to remind them of what was available. This rule was imposed to avoid the possibility of subjects simply naming off the parts in order and marking the available part types so that the list became a surrogate for the information normally available from looking at the parts bin.

6.5.3 Restricted Simulation (RS) Trials

Under this condition, the opportunities for Group 2 subjects to gather information about constraints was reduced. These subjects were only allowed five "simulations". Each time a subject used the "start machine" feature, the experimenter would inform the subject of how many of their allotted simulations they had used and how many they had remaining. If a subject used all of his simulations without solving the problem, he would be allowed to continue adding, deleting, and manipulating parts within the workspace, but would not be allowed to use the simulation feature. If a subject wanted to test whether or not his current configuration solved the problem, he was allowed to ask the experimenter to run the machine for him. The subject would be asked to turn his head while the experimenter ran the machine. The experimenter would then inform the subject whether or not the problem had been solved. Before the subject was allowed to view the screen again, the experimenter would reset the machine (by clicking the mouse button) to the configuration the subject had constructed before the simulation attempt.

6.5.4 Increased Parts (IP) Trials

Trials performed under this condition were based on problems from the original set provided in TEMIM, but were modified with respect to the number and/or types of parts available to solve it. For Group 1 subjects, the number and/or types of parts included in the original TEMIM problem were reduced. There were still enough parts to enable subjects to construct (at least) one solution, but the number of potential arrangements of objects with which subjects could attempt a solution was more highly constrained than in the original scenario. Group 2 subjects were presented with a slightly increased number of parts and part types. The additional parts were chosen carefully such that they would not allow for a trivial solution to the problem. Care was taken to choose parts that served largely as distractors and did not significantly increase the number of additional possible

solutions, although it was quite difficult to be certain in all cases that this was avoided. Appendix B details the parts available in all trials.

6.6 PROCEDURE

The experiment consisted of 5 sessions, not including the initial screening test for spatial ability. Sessions were performed on separate, consecutive days (if possible). In the introductory session, the experimental protocol was explained to subjects and they were presented with a series of tutorial problems to solve. The next four sessions each consisted of four data collection trials. A "recipe" test (to be described below) was administered before the tutorial problems, before the first data collection session, and after the last data collection session. Table 6 presents a summary of the experimental protocol.

TABLE 6. Summary of Experimental Protocol.

SESSION	TRIALS / TESTS	GROUP 1 CONDITION	GROUP 2 CONDITION
Screening	Spatial Ability Test		
Session 1	Recipe Test		
	Tutorial Problems (21)	N	N
Session 2	Recipe Test		
	Trial 1	N	N
	Trial 2	N	RS
	Trial 3	IP	IP
	Trial 4	N	RS
Session 3	Trial 5	N	RS
	Trial 6	N	CB
	Trial 7	N	N
	Trial 8	N	CB
Session 4	Trial 9	IP	IP
	Trial 10	N	N
	Trial 11	IP	IP
	Trial 12	N	N
Session 5	Trial 13	IP	IP
	Trial 14	N	RS
	Trial 15	N	CB
	Trial 16	N	CB
	Recipe Test		
	Debriefing		

6.6.1 Spatial Ability Test

During the screening session subjects were asked to complete the VZ-2 paper folding test of spatial visualization ability (Ekstrom et al., 1976; see section 4.0.5 for a detailed description). Subjects scoring lower than 14.0 were assigned to the low spatial ability group. Subjects scoring 18.0 or above were assigned to the high spatial ability group. These cutoff points were based on those used by Vicente et al. (1987, 1988). Subjects achieving scores outside of the acceptable ranges were not included in the remainder of the study but were paid $7 for their participation. Subjects falling into either the high or low spatial categories were assigned randomly to either Group 1 or Group 2.

6.6.2 Introductory Session

The introductory session was identical for all subjects. Subjects were first asked to fill out a demographic questionnaire. They were then given a set of instructions for the experiment, followed by a demonstration of the basic use of TEMIM. In order to ensure that subjects did not simply give up during difficult problems, they were informed that all of the problems presented had solutions. Having learned the basic objectives and rudimentary rules of TEMIM, they were then asked to complete the "recipe test" (see section 6.7). They were then given a brief set of instructions regarding the collection of verbal protocols before beginning the tutorial problems. Copies of all forms and instructions used in this session can be found in Appendix A.

6.6.2.1 Tutorial problems. The TEMIM software package includes a series of 21 "tutorial" problems which introduce subjects to the game. These problems allow subjects to become familiar with the rules of the game as well as with the use of the various parts. All of the part types encountered over the remainder of the experiment were introduced to subjects during the trials in this session. Subjects were required to complete all 21 of these problems, although based on the results of the pilot study, a 10 minute time limit was enforced for each trial. Because the object of these problems was to teach subjects the basics of using TEMIM, subjects who exceeded the time limit for a particular trial were shown a solution to the problem by the experimenter. The experimenter would arrange the parts into a configuration constituting a solution and allow the subjects to view the solution being "simulated" (the same solutions were always revealed for individual problems).

6.6.2.2 Data Collection. A second purpose of the introductory session was to allow subjects to become comfortable with the use of a video camera and microphone for data collection. During all trials, the computer monitor was videotaped and subjects' voices were recorded. This allowed for subjects' actions within the game to be recorded (e.g., part placement, simulation, etc.) and provided a full recording of all verbalizations.

Subjects were encouraged to verbalize their thought processes as fully as possible. If subjects fell silent for long periods, the experimenter would prompt them by saying "Please keep talking", or "Don't forget to talk". The phrase "What are you thinking?" was specifically avoided as per the recommendations of Ericsson and Simon (1993).

6.6.3 Data Collection Sessions

Trials during the data collection sessions were conducted in a similar fashion to the tutorial problems. The time limit for these problems was 30 minutes. No solutions were shown to subjects failing to complete problems. Data collection was performed in the same manner as described above.

6.6.4 Recipe Tests

(For a detailed description of this test, please refer to section 6.7). The recipe test was given immediately after TEMIM was introduced, so that subjects would have a minimal though sufficient understanding of the nature of the problems to complete the test. Before beginning trials in the second session, subjects were given the recipe test once again. At this point the subjects had become somewhat familiar with TEMIM, having completed the 21 tutorial problems. However, Group 1 and Group 2 had not yet begun to perform trials under different conditions. Therefore this point was viewed as a potentially informative one at which to perform these tests a second time. The test was administered for a third and final time after trials had been completed in the final session

6.6.5 Debriefing

The final procedure in the experiment was an informal interview with subjects about their approaches to problems in TEMIM. The same questions were asked of subjects in both groups with the exception of a pair of questions asked of subjects in group 2 which related to the trials performed under reduced constraint conditions. The list of questions for each group is presented in Appendix C. Results were not available at the time of the debriefing to inform subjects about their performance, but the subjects were encouraged to return to inquire about the results later on if they so desired.

PERFORMANCE MEASURES

The primary sources of data in this experiment were the time-stamped logs of actions and verbalizations, both extracted from the video recordings and accompanying

verbal protocols collected during all trials. Creation of the data logs was performed using the Timelines video analysis system (Harrison, Owen, and Baecker, 1994). The recipe tests served as secondary data sources. The performance measures derived from each of these sources will now be described.

6.7 1 Completion Times

The only product measure used in this study is trial completion time. The time taken to complete each trial was extracted from the data logs and recorded. Unfinished trials were assigned a completion time of 30 minutes, the maximum time allowed.

6.7.2 Action Protocols

Four categories of actions were coded. Firstly, all instances of subjects' use of the simulation feature were recorded. Second, a recording of each time a part was removed from the parts bin and added to the workspace was made. Third, all instances of parts being deleted from the workspace and put back in the parts bin were recorded. Finally, all occurrences of subjects' modification of parts within the workspace were recorded. This category included all cases of subjects moving or modifying (e.g., re-orienting, or otherwise adjusting) any of the parts in the workspace.

6.7.3 Verbal Protocols

Two categories of verbalizations were coded: abstract references and goal definition statements. Each of these is outlined below.

6.7.3.1 Abstract References. These are statements referring to abstract concepts such as energy, power, force, motion, and "action". Such statements indicate subjects' consideration of generic affordances of objects in terms of abstract concepts. The frequency of such statements is taken as an indication of the level of abstraction at which subjects frame the problem and perceive the objects in the problem. This category of statements is based loosely on Rasmussen's (1986) definition of the Abstract Function level in his Abstraction Hierarchy (AH). Only statements which referred to these concepts as generic phenomena were coded in this category. Thus, statements such as "I need to transfer this motion over here", were included, but statements like "I need to get this thing moving" were not. These latter types of statement were often ambiguous regarding the level of abstraction of subjects' reasoning and were usually tied specifically to concrete elements of the problem.

6.7.3.2 Goal Definition Statements. These are statements where subjects explicitly refer to either a subgoal or to the overall goal. They are typically prefaced with phrases

such as "I need to...", "I have to...", "I want to...", "how can I...", etc. These goal statements must refer to genuine "problems" according to Duncker's (1945) definition, which stipulates that the solution not be immediately obvious to the subject and not be simply an exercise in retrieving the entire solution from memory. In other words, a goal statement is taken to be an identification of a need, for which the subject is forced to actively consider (and possibly attempt to resolve) the question of how it is to be satisfied. Statements referring to why an action was taken after the fact are not included in this category since they are ambiguous as to whether the goal of the action was covertly considered by the subject beforehand or whether they represent a post-hoc justification of a bottom-up, data-driven action. The frequency of such statements is intended to indirectly measure the extent to which subjects employ a goal-driven, backward chaining type of reasoning. Statements which indicate blindly associative strategies, such as "what can I use this object for?", or "what objects go together with these objects?" are not coded in this category because they do not imply a specific causal goal.

6.7.4 Recipe Test

This test was adapted from a measure used by Irmer and Reason (1991). Subjects were asked to write out a set of generic instructions for solving problems in TEMIM. They were told to make the instructions detailed enough that somebody who had never used TEMIM before could understand and use them. Completion times were recorded but once again, subjects were told that finishing quickly was not the object of the test. A copy of the instructions provided to subjects for this test can be found in Appendix D.

7. RESULTS

The protocols for all 16 data collection trials of each of the 16 subjects were analyzed, translating into approximately 40 hours of video in total. The results will be presented in four parts. First, the results of ANOVAs conducted on the product and process measures will be reported. Second, results of an analysis of correlations between variables of interest will be presented. Third, results from the recipe tests will be presented. Finally, a brief discussion of some of the informal observations made during the experiment which were not captured by the analyses will be presented.

7.1 ANOVAs

A 2x2x4x4 mixed ANOVA was conducted for each of the process and product measures. The independent variables are listed below in Table 7.

TABLE 7. Independent variables

Independent Variable	Levels
Spatial ability	[high, low]
Constraint group	[1, 2]
Problem type	[CB, IP, N, RS]
Problem number[3]	[1, 2, 3, 4]

The dependent variables consist of 1 product measure and 3 process measures, which are listed below in Table 8. All of the process measures were calculated as rates over time.

TABLE 8. Dependent variables

Dependent variable
Completion time (product)
Simulation rate (process)
Abstract reference rate (process)
Goal definition rate (process)

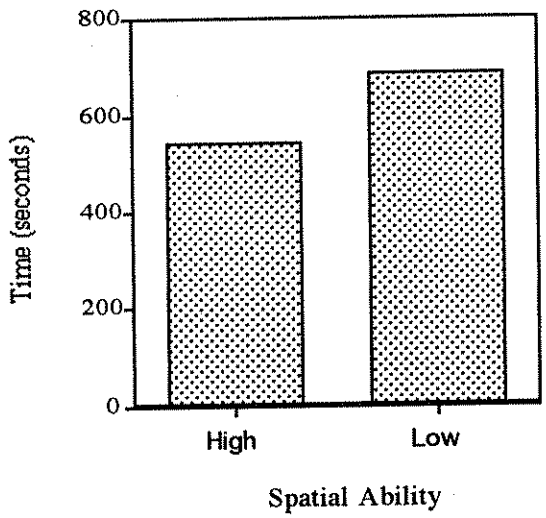
[3] Problem number *i* refers to the *i*th appearance of a given problem type.

The results will be organized by dependent variable. Because no reliable method was available for classifying problems, effects involving problem number were generally not investigated. These effects were felt in general to represent the results of differences between individual problems, which were not of interest in the current study.

7.1.1 Completion time

The results for completion time revealed a significant main effect for spatial ability [$F(1,12)=5.14$, $p<0.05$]. This effect is graphed in Figure 3.

FIGURE 3. Average Trial Completion Time by Spatial Ability Group



The difference in average completion times between the high and low spatial groups was on the order of two and a half minutes (541 vs. 686 seconds), a substantial difference when viewed against the overall average completion time of approximately ten minutes (614 seconds). The prediction that spatial ability might act as a significant performance shaping factor thus seems to be substantiated.

The difference between the constraint groups is even more pronounced [$F(1,12)=15.25$, $p<0.01$], with the reduced constraint group (Group 2) averaging 738 seconds while the baseline group (Group 1) averaged 489 seconds; an average difference of over four minutes (see Figure 4). However, there is also an interaction involving constraint group and problem type [$F(3,36)=6.20$, $p<0.01$]. This effect, graphed in Figure 5, reveals that for Normal trials, Group 2 actually had a faster average completion time than

Group 1 (542 seconds vs. 598 seconds). This contrasts sharply with the differences revealed for all of the other trial types. Indeed, t-tests (unpaired) for simple effects show that group 1 was faster for CB Trials [$t(62)=3.584$, $p<0.001$, one-tailed], for IP trials [$t(62)=3.715$, $p<0.001$, one-tailed], and for RS trials [$t(62)=2.876$, $p<0.01$, one-tailed], while there was no significant difference for Normal trials [$t(62)=1.076$, n.s.].

FIGURE 4. Average Trial Completion Time by Constraint group

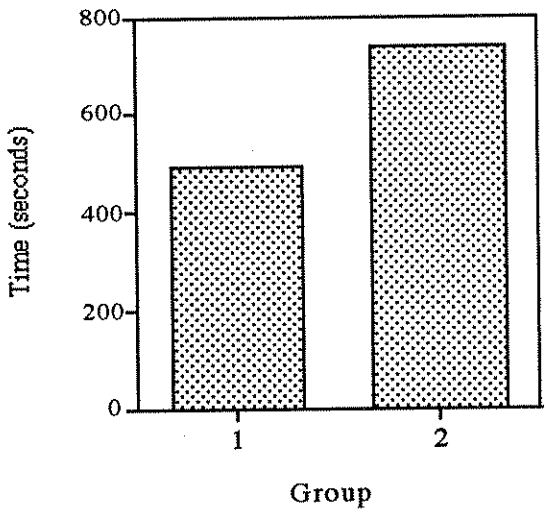
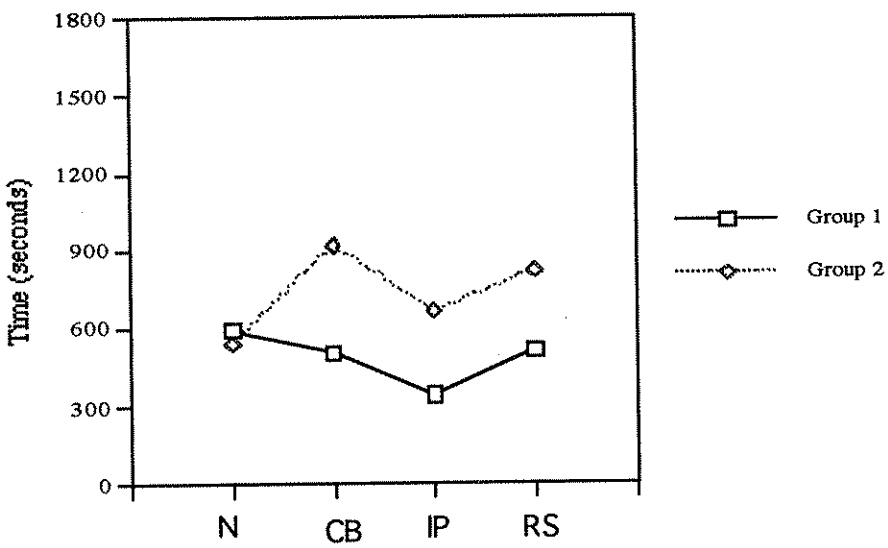
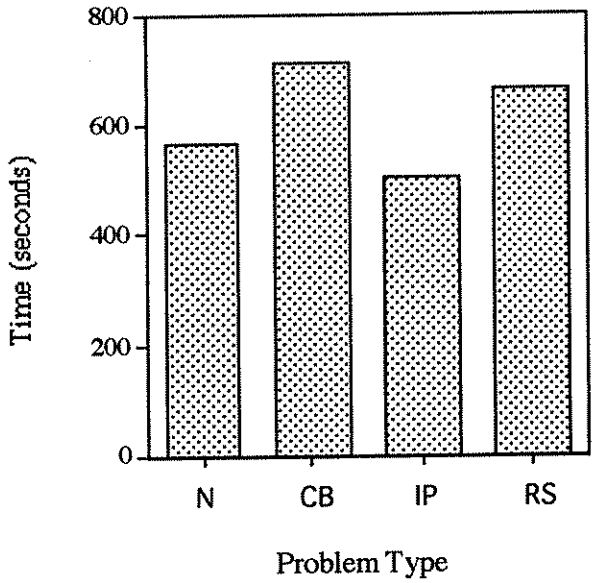


FIGURE 5. Interaction of Constraint group and Problem Type on Trial Completion Time



A main effect was also found for problem type [$F(3,36)=5.07$, $p<0.01$] (see Figure 6). Pairwise comparisons were performed using a Student-Newman-Keuls test with $\alpha=0.05$, and revealed that IP trials were performed significantly faster than both CB trials and RS trials, which were not significantly different from one another. Normal trials were not significantly different from any of the other trial types.

FIGURE 6. Average Trial Completion Time by Problem Type



This effect may indicate some combination of effects due to the differences between the individual problems which were used under each of the four trial types, and to the differences between the trial types themselves.

A significant main effect for problem number was found [$F(3,36)=18.04$, $p<0.0001$], as was an interaction between problem type and problem number [$F(9,108)=7.49$, $p<0.0001$]. These results indicate that certain problems were completed more quickly than others However, as indicated above, there is no clear way to interpret these results given the lack of a relevant method for classifying problems. These effects were therefore not investigated further.

7.1.2 Simulation Rate

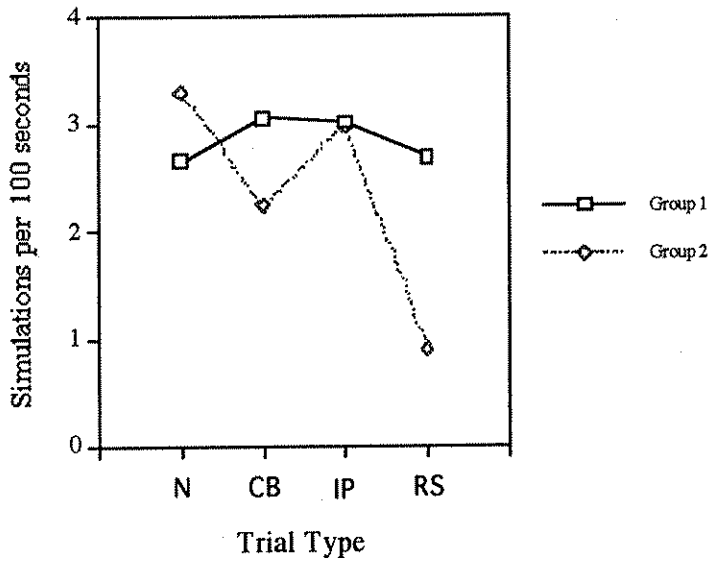
The rate of subjects' use of the simulation feature (measured in simulations performed per 100 seconds) was intended to act as a gauge of the frequency with which

subjects sought new information about the kinematic and dynamic relations between objects and systems of objects. Recall that, because of the apparently highly spatial nature of the experimental task (see section 5.1), it was expected that subjects in the low spatial ability group might, in general, use the simulation feature more than the high spatial group in order to compensate for their presumably reduced ability to infer the kinematics and dynamics of a given configuration through mental simulation alone. This prediction was however not supported by the data, which revealed a non-significant difference between the spatial ability groups [$F(1,12)=0.70$, n.s.]. The high spatial group had an average simulation rate of 2.81 simulations per 100 seconds, while for the low spatial group the figure was slightly lower at 2.43.

The data revealed a significant interaction between constraint group and problem type [$F(3,36)=13.27$, $p<0.0001$], which is displayed graphically in Figure 7. Simple effects were tested for using t-tests (unpaired) at each level of problem type. These revealed a highly significant difference between the constraint groups' use of the simulation feature for RS trials [$t(62)=5.78$, $p<0.001$, one-tailed], with subjects in Group 1 averaging almost three times the rate of simulations of subjects in Group 2 (2.70 vs. 0.91 simulations per 100 seconds). Significant simple effects were also detected for CB trials [$t(62)= 2.38$, $p<0.05$, one-tailed] where again, subjects in Group 2 used the simulation feature less than subjects in Group 1 (2.26 vs. 3.06 simulations per 100 seconds, respectively). For Normal trials this trend was reversed, with subjects in Group 2 displaying a higher rate of simulation the subjects in Group 1 (3.32 vs. 2.67 simulations per 100 seconds). However, no simple effect was detected [$t(62)=-1.49$, n.s.]. No simple effect was present for IP trials [$t(62)=0.14$, n.s.], with the groups displaying very similar rates of simulation (3.03 and 2.98 simulations per 100 seconds for Group 1 and 2 respectively).

Not surprisingly, the results also showed a significant main effect for problem type [$F(3,36)=15.57$, $p<0.0001$]. Pairwise comparisons between problem types were performed using a Student-Newman-Keuls test with $\alpha=0.05$, which showed that RS trials were performed with a significantly lower rate of simulation than all other trial types. None of the remaining trial types differed significantly. Significant effects were also observed for problem number [$F(3,36)=20.11$, $p<0.0001$], for the interaction between problem type and problem number [$F(9,108)=7.27$, $p<0.0001$], and for the three way interaction of constraint group, problem type, and problem number [$F(9,108)=3.09$, $p<0.005$]. Again, these effects involving problem number were felt to be primarily due to differences between individual problems and were therefore not considered further.

FIGURE 7. Interaction of Constraint group and Problem Type on Rate of Simulation



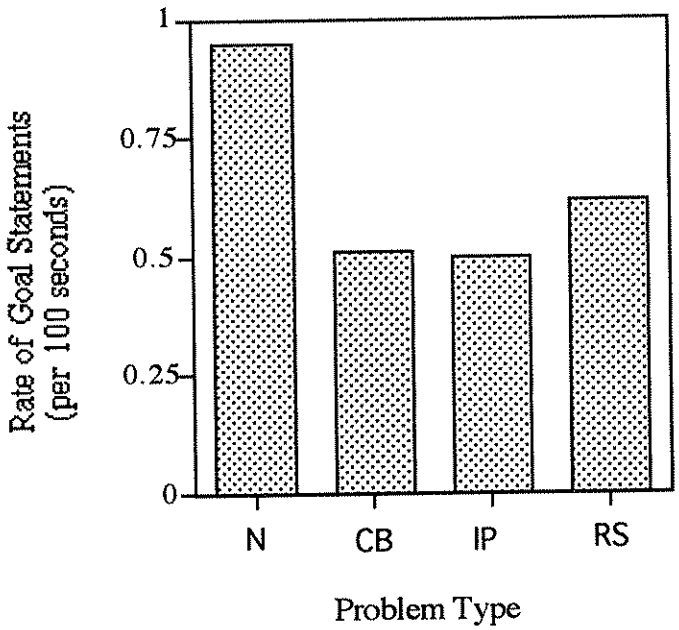
7.1.3 Goal Definition Rate

Recall that the rate of goal definitions was intended to act as an indirect measure of subjects' tendencies to employ an explicitly goal-oriented, backward-chaining style of reasoning during problems. The results revealed no significant differences between spatial ability groups [$F(1,12)=3.24$, n.s.]), or between constraint groups [$F(1,12)=0.34$, n.s.]. The means do however reveal a fairly wide difference among the spatial ability groups: 0.80 goal definitions per 100 seconds for the high spatial group vs. 0.49 for the low spatial group. The constraint groups are more similar, with Group 1 displaying an average rate of goal definitions of 0.70 per 100 seconds against Group 2's rate of 0.60 per 100 seconds. There was a significant main effect for problem type [$F(3,36)=8.33$, $p<0.001$], which is shown in Figure 8. Pairwise comparisons were performed using a Student-Newman-Keuls test with $\alpha=0.05$. This revealed that subjects' average rate of goal definitions was significantly higher on Normal trials compared with the remaining trial types. None of the other trial types differed significantly. No effects involving problem number were detected.

7.1.4 Rate of Abstract References

The rate at which subjects made reference to concepts such as energy, mass, and force was used to gauge subjects' tendency to represent problems abstractly, in terms of a relatively small set of generic affordances corresponding roughly to the Abstract Function level in Rasmussen's (1985) Abstraction Hierarchy. The results of the ANOVA failed to

FIGURE 8. Effect of Problem Type on Rate of Goal Definitions

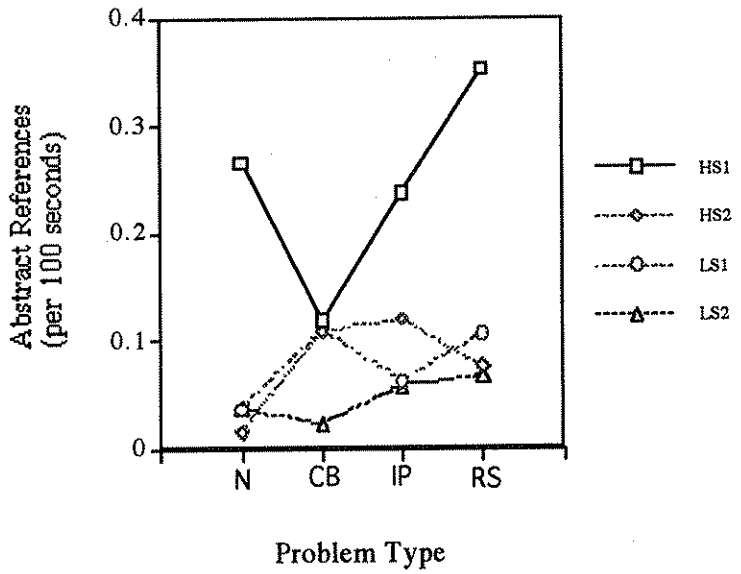


show any differences between spatial ability groups [F(1,12)=2.39, n.s.] or between constraint groups [F(1,12)=2.36, n.s.]. The means for the spatial ability groups were nonetheless highly separated, with the high spatial group averaging 0.16 abstract references per 100 seconds, against the low spatial group's average of 0.06. The difference in means is identical between the constraint groups. Group 1 displayed an average rate of 0.16 vs. Group 2's rate of 0.06. Some investigation reveals that these differences can perhaps be attributed to individual differences. The top subject on this measure, SJ of the high spatial Group 1 subjects, displayed an average rate of 0.54 abstract references per 100 seconds, which was almost twice that of the next subject (CS), who was also in the high spatial half of Group 1. Remarkably, SJ's rate was nearly 100 times the rate of the subjects who ranked lowest on this measure. Such a wide spread in values for this measure would seem to point towards individual differences as a source of variability. It should also be noted that these rates are in fact very low, corresponding to less than one such reference per trial for most subjects, indicating that abstract verbalizations were not, in general, heavily used.

A significant effect was detected for the interaction between spatial ability, constraint group, and problem type [$F(3,36)=3.43$, $p<0.05$]. This effect is shown in Figure 9 and reveals that high spatial subjects in Group 1 displayed much higher rates than any of the other groups for all trial types except CB trials.

A significant effect was also detected for the interaction of problem type and problem number [F(9,108)=4.07, p<0.0005], indicating that there were significant differences between individual problems. However, this effect was not investigated further.

FIGURE 9. Interaction of Spatial Ability, Constraint group, and Problem Type on Rate of Abstract References



7.1.5 Discussion

The results of the ANOVA clearly reveal that the two primary independent factors in the experiment (spatial ability and constraint condition) did indeed have significant impacts on performance. The discussion will begin with the results pertaining to spatial ability, followed by the results pertaining to constraint group, followed briefly by those related to problem type.

7.1.5.1 Spatial ability. The comparison between the spatial ability groups is unambiguous - high spatial subjects clearly performed better than low spatial subjects. The results did not, however, provide any clues about the precise origin of the differences in performance since spatial ability did not appear in any other significant main effects or interactions (for completion time or for any of the process measures), with the single exception of the three way interaction between spatial ability, treatment condition, and problem type on the rate of abstract references. High spatial ability thus seemed to lead to a

general superiority in performance, which was not the result of an advantage on any particular problem type. The results suggest that there was no simple relation between spatial ability and any of the process measures examined which could account for the high spatial subjects' generally superior performance (although the apparently wide differences in means between the spatial groups on some measures is at least suggestive of some of the sources). Thus, the difference in performance between the spatial ability groups may have been due to a more complex relationship arising from the combination of non-significant differences on the process measures, or (at least partially) due to differences in some other mediating factor or process which was not directly captured by the measures used in the present study. This question will be explored further in the general discussion.

The lack of any significant effects involving spatial ability for the process measures deserves some discussion. In particular, the lack of any such effects with respect to the rate of simulation seems especially notable. Based on the results of the pilot study, it had seemed reasonable to expect that the difference in spatial ability between the two groups would have somehow manifested itself in the rate of simulation. Specifically, it had been thought that subjects in the high spatial group would have less need to use the simulation feature to gain information about the kinematics and dynamics of system configurations, being more readily able to generate this information mentally. However, since the simulation feature was a relatively cheap source of unambiguous feedback (with the exception of the case of RS trials for Group 2 subjects where simulating had a very definite cost associated with it), it is very possible that <u>all</u> subjects favoured using the simulation feature over deliberate mental simulation, the products of which were naturally approximate at best, and potentially involved significantly greater cognitive effort. That being said however, it is possible that the high spatial subjects were in fact able to utilize the information provided by the simulation feature more fully or effectively than low spatial subjects, essentially extracting more information from the same overall rate of simulation. This idea will be pursued in subsequent sections.

Another factor which would serve to obscure differences between the groups in terms of simulation rates comes from an informal observation made regarding how subjects used the simulation feature. For most subjects, the bulk of the use of the simulation feature seemed to occur during episodes of what might be termed "fine tuning" of system configurations. Precise part placement was often a critical factor in solving problems in TEMIM, a fact which most subjects seemed to discover rather quickly. The exact placement of an object relative to other objects, down to the finest grain of precision possible, could sometimes significantly affect the progression of a given causal chain. Because of this, subjects often seemed to be willing to invest considerable time in making

micro-adjustments to individual parts in order to elicit the desired behaviour from the system, using an iterative sequence of adjustment and simulation. This was especially true in cases where several user-configurable parts were involved. Because it was the combination of the positions of the objects that was usually the critical factor in eliciting a given higher-order function from the group of elements, subjects were able to make a combinatorially larger number of adjustments to the system in trying to achieve the desired behaviour. In these sorts of scenarios, spatial ability was not likely a factor, the reason being that at this level, the properties of the objects were often sufficiently arbitrary that the only possible way to gain any productive information about the behaviour of the configuration was to run the machine.

It does not seem unreasonable to assume that these "fine tuning" episodes were experienced equally by both spatial ability groups, since they seemed to be associated with the arbitrary properties of TEMIM, relatively independent of other factors). Therefore, if the claim that these episodes are where the largest proportion of simulation occurred is approximately correct, then it is perhaps not surprising that the rate of simulation for the spatial ability groups were in fact similar. Any other differences in the rates of use of the simulation feature, which may or may not have been associated with strategies mediated by spatial ability, would have been dominated by the overall similarity with respect to fine tuning episodes.

The main implication of this interpretation is that the rate of simulation was not as directly sensitive to differences in strategy which may have existed between the spatial ability groups. It is certain that the simulation feature was used for other purposes than "fine tuning", purposes for which spatial ability may have been influential, but the present analysis cannot differentiate among these. The differences in spatial ability between the groups may also have manifested themselves at an entirely different level of behaviour, which was not captured by any of the process measures in this study.

7.1.5.2 Constraint group. The difference in performance between constraint groups is also quite clear in the results. The interaction between constraint group and problem type shows that for Normal trials, where both groups performed under identical conditions, the groups' performance was not significantly different (in fact, Group 2 performed slightly better than Group 1). However, for each of the other three types of trials where Group 2 performed under reduced constraint conditions, there is a significant performance difference, with Group 1 consistently outperforming Group 2. This seems to confirm that the performance differences are in fact due to the constraint reducing manipulations and not to some general superiority on the part of Group 1.

Interestingly, in a situation similar to that observed for spatial ability, there were no other main effects of constraint group on any of the process measures. However, constraint group was involved in a significant interaction with problem type on the rate of simulation. This effect revealed, not surprisingly, that for RS trials (where Group 2 was limited to five regular uses of the simulation feature), Group 2's average rate of simulation was significantly less than that for Group 1. Naturally, Group 2 subjects were more selective about when to use the simulation feature for these trials. Also, for subjects who exceeded the limit of 5 simulations and were no longer allowed to observe simulation runs, the utility of the simulation feature became much reduced, thus leading to less incentive to use it as a source of feedback. Interestingly, at least two subjects in Group 2 (DN and TB) found a way to extract information from these simulation runs in spite of not being able to observe them. These subjects noted that the amount of time which elapsed from the beginning of a particular simulation to the point where the experimenter informed them that their configuration had not achieved the goal often changed noticeably for different configurations. Under the assumption that longer time periods indicated that the causal chain(s) they were attempting to create were successfully proceeding further, they used the length of this time period to estimate approximately where in the intended causal chain their configuration was deficient. It is possible that other Group 2 subjects were also using this highly adaptive strategy. However, only these two subjects made verbal comments revealing that they were in fact doing so.

Group 2's lower rate of simulation for CB trials can likely be attributed to the fact that, for these trials, Group 2 subjects spent a relatively large amount of time concerned with discovering what parts were actually available to use in attempting a solution. Because subjects could not anticipate by scanning the available parts, any solution concept was susceptible to turning into a dead end if a critical type of part was found to be unavailable. This may have meant that subjects were less likely to invest the time in fine-tuning a partial solution without having a clearer idea of the remaining parts available (recall that fine-tuning was often observed to be where the heaviest use of the simulation feature took place). The lack of differences between the groups' simulation rates on Normal and IP trials can likely be attributed to the fact that both groups performed these trials under what were in fact very similar conditions, the only exception being the number of parts available to Group 2 subjects in IP trials.

7.1.5.3 Problem type. Problem type was involved in a number of the interactions described above but was also found to reveal a main effect on completion time, rate of simulation, and goal definition rate. The effects on completion time and simulation rates can be tied to the interactions described above, but the effect on the rate of goal definitions

cannot. The fact that Normal trials seemed to lead to a significantly higher rate of goal definitions than any of the reduced constraint trial types can be attributed either to differences in the problems used for each trial type, or perhaps to differences between the constraint groups' performance which were not revealed by any of the other measures.

7.2 CORRELATIONAL ANALYSIS

In order to attempt to identify relations among the dependent variables and to try to determine if these relationships differed across levels of the major factors of interest, a series of correlational analyses were performed. The questions which these analyses attempted to answer were: which, if any, process measures are associated with good performance? Do these associations differ across levels of spatial ability or constraint? Does the relationship between spatial ability and the process measures differ between the constraint groups?

Because the assumption of sampling from a bivariate normal distribution was felt to be unrealistic for many of the comparisons performed and because the relations between variables were not necessarily expected to be linear, Spearman rank-order correlation coefficients (r_S) were used rather than Pearson product-moment correlation coefficients. All 16 trials performed by each subject were used as data, resulting in 256 total data points.

7.2.1 Product versus Process

This first analysis involved the product variable (completion time) and the set of process variables used in the ANOVAs. Correlations were calculated and compared across each spatial ability group and each constraint group, with n=128 for each correlation. The results are shown in Table 9 below. The cells contain r_S values and the associated p-values.

7.2.2 Spatial Score vs. Product and Process

This analysis sought to discover if the actual score which subjects achieved on the ETS VZ-2 test for Spatial Visualization was significantly related to any of the process measures, or to trial completion time. Spearman rank order correlation coefficients were used once again because the spatial ability scores were clearly not distributed normally, neither within spatial ability groups nor overall.

TABLE 9. Correlations of Process variables versus Trial Completion Time (shaded cells denote $p<0.05$)

Sample Group	Simulation Rate	Rate of Goal Definitions	Rate of Abstract References
Group 1	-0.059, 0.5105	0.054, 0.5465	0.165, 0.0634
Group 2	-0.194, 0.0280	-0.160, 0.0707	0.322, 0.0002
High Spatial	-0.218, 0.0134	-0.162, 0.0676	0.082, 0.3597
Low Spatial	-0.116, 0.1926	0.081, 0.3611	0.346, 0.0001

The scores achieved by all of the subjects on the spatial ability screening test can be found in Table 5. Correlation coefficients were calculated within each constraint group with all trials used as data points, giving a total of 128 data points for each analysis. The results are shown in Table 10 (cells include r_S values and p values).

TABLE 10. Correlations of Spatial Ability Score with Product and Process Measures by Constraint group (shaded cells denote $p<0.05$).

Constraint group	Trial time	Simulation Rate	Rate of Goal Definitions	Rate of Abstract References
1	-0.109, 0.2193	0.245, 0.0053	0.274, 0.0018	0.280, 0.0014
2	-0.170, 0.0550	0.008, 0.9327	0.152, 0.0873	-0.030, 0.7332

7.2.3 Discussion

The results will be discussed in the same order that the analyses are presented above.

7.2.3.1 Product vs. Process. The negative correlations between the rate of simulation and performance displayed by all groups were in the direction expected, suggesting that there may indeed be a relationship between the rate of simulation and performance. The fact that the relationship is significant for the high spatial group but not

for the low spatial group is quite curious. The original expectation had been that the simulation feature would be more valuable to low spatial subjects as a means of compensating for their presumably reduced ability to mentally infer the kinematics and dynamics of object configurations. However, these results suggest that it was in fact the high spatial subjects who benefited more, relative to one another, from increased use of the simulation feature.

One possible interpretation for this result is that, as alluded to in the discussion of the ANOVA results, high spatial subjects may have been able to more fully utilize the information provided by the simulation feature, not only as a test of current configurations, but also, for instance, as a way to enable predictions or to generate ideas about possible future arrangements of objects. High spatial subjects who used the simulation feature more extensively may thus have been able to gain an advantage over their fellow subjects because they were, in general, able to extract a greater amount of useful information by doing so. For low spatial subjects on the other hand, if it is assumed that they were not as readily able to take advantage of information available through the simulation feature (beyond the sufficiency of current configurations), then they would not be expected to be as strongly differentiated with respect to performance as a result of the use of simulation.

Group 2's significant negative correlation between simulation rate and trial completion time can likely be attributed to the specific effects of two of the reduced-constraint trial types. First, during RS trials, the Group 2 subjects who performed well were typically those who were able to solve the problem within the limit of 5 regular uses of the simulation feature. Subjects who did not tended to have a great deal of difficulty in completing the trials, leading to very lengthy completion times during which subjects had relatively little incentive to simulate, because they were not allowed to observe the monitor while the simulation was in progress. Therefore, subjects who exceeded the 5 simulation limit tended to take a long time to complete these trials and did not tend to simulate particularly frequently, thus leading to a lower overall rate of simulation. On CB trials, subjects who were able to quickly identify and retrieve the parts relevant for the problem tended to begin fine tuning arrangements of objects much more quickly than subjects who had difficulty in identifying the available and/or relevant parts (and therefore generally had less to gain by simulating frequently). Subjects who exhibited the former type of performance also naturally tended to complete the trials more quickly. Thus, a higher rate of simulation was also associated (though not necessarily in any causal way) with good performance.

The results suggest (weakly) that the rate of goal definitions may have been more influential on performance for the high spatial subjects than the low spatial subjects. The

negative correlation displayed by the high spatial group, though only marginally significant, is provocative when taken against the positive correlation of the low spatial group. High spatial subjects who set goals more frequently tended to be those who performed better. While the direction of causality cannot be determined, it is possible to speculate that high spatial subjects were able to set more realistic or more productive goals based on better conceptions of the implications of their choices. Therefore, high spatial subjects who set goals frequently tended to benefit from the practice relative to other high spatial subjects who did not tend to set goals as frequently. The fact that low spatial subjects did not show a similar relation may be due to the fact that the goals they set were based on relatively ill-conceived notions of the configurations required to solve the problem at hand.

The fact that a similar set of results were obtained between the constraint groups is perhaps not as surprising. The marginally significant negative relation for Group 2 may be interpreted as suggesting the increased importance of setting goals under reduced constraint conditions. The problems faced by Group 1 may have been sufficiently well-constrained so as to make goal definitions less valuable relative to other methods of approaching problems.

The pattern of results with respect to the rate of abstract references is very surprising. Firstly, correlations for all groups are positive, suggesting that subjects who made more abstract references were in fact those who performed relatively poorly within their groups. The original expectation had been that subjects who tended to represent problems and objects abstractly would gain a substantial advantage by doing so. Abstract references are presumed to indicate that subjects are representing the problem in terms of abstract functions which serves to aid in understanding and resolving gaps in the causal structure by providing some initial constraint on the types of objects or systems of objects which will satisfy the current objectives within the problem. These results suggest that there is no such relation between abstract representations and performance, particularly for the low spatial and Group 2 subjects.

7.2.3.2 Spatial score vs. Product and Process. With respect to the product measure, both constraint groups show negative correlations between spatial score and trial completion time, though neither is significant. This is consistent with the results of the ANOVA which revealed no significant interaction between constraint group and spatial ability group with respect to trial completion time. The direction of the correlations is however consistent with expectation in that higher spatial ability seems to be associated with lower trial completion times.

The results for the process measures are immediately striking in that several of the process measures are correlated at highly significant levels with spatial ability score for Group 1, whereas none of the measures reveal significant correlations with spatial score for Group 2. Individually, the results are difficult to interpret, but as a whole they suggest that spatial ability was a powerful mediating factor in many of the process measures. The fact that no significant relations were found for Group 2 can be taken to indicate that the reduced constraint conditions acted in such a way as to significantly attenuate any effects of spatial ability on the process measures collected.

7.3 RECIPE TESTS

The recipe tests proved to be very difficult to code. Recall that the recipe test asked subjects to write out a general set of instructions for solving problems in TEMIM which could be used by someone who had never seen TEMIM before. Because of the wide variety in the problems subjects were asked to solve, and because the instructions specified that they try to list a general procedure for solving problems, it seemed to be difficult for most subjects to be very concrete. The types of recipes therefore differed widely across subjects, with some subjects presenting relatively rigid algorithms for solving problems, while other subjects favoured declarative descriptions of the characteristics of TEMIM, with relatively little emphasis on direct instructions for action. Most subjects fell somewhere in between, interspersing procedural instructions with declarative hints, caveats, or other general information pertaining to TEMIM. Some subjects gave one general method while others listed several heuristics. Given this wide variation not only in styles of completing the recipe tests, but in the methods and heuristics listed themselves, classification and interpretation of the responses was difficult.

Only the results for the subjects' final recipe test will be presented here. This test was completed after subjects had performed all trials, at the end of the final session. The recipes subjects gave during this session were therefore presumed to represent a relatively stable picture of the way they believed problems should be approached. The analysis consisted of a count of the number of subjects making references to specific heuristics for solving problems in TEMIM. The set of heuristics is not necessarily representative of the entire set of heuristics which subjects employed during the trials, but does capture most of the methods explicitly mentioned in the recipe tests. Also recorded in this analysis were counts of metastrategic statements, as well as statements of declarative hints, caveats and general information about TEMIM (see Table 11 below). The specific categories are:

1. Means-ends strategies. This category was somewhat unusual in that it was most typically implied through a number of statements rather than stated explicitly and coherently. If subjects included statements to the effect of observation and evaluation of the difference between the current state and the goal state, as well as statements referring to the selection of operations to reduce those differences, then they were given credit for using this heuristic in their recipe.

2. Problem partitioning. Subjects were given credit for reference to this heuristic if they specifically made mention of a strategy of dividing the problem into subproblems or sections.

3. Associative strategies. This heuristic was attributed to subjects whose recipes specifically mentioned the "matching" of pairs or groups of objects based on past experience of typically paired or grouped objects.

4. Backward chaining. This strategy was one of "working backwards from the goal". In other words, this strategy referred to the construction of a causal chain by backward chaining starting at the goal and working towards some (possibly undefined) initiating object or event.

5. Forward chaining. This strategy was the complement of subgoaling, namely a forward causal chaining strategy. Subjects who referred to a method of beginning with some initiating object or event and attempting to extend the causal chain forward towards a (possibly undefined) goal were given credit for reference to this heuristic.

6. Metastrategic statements. These were statements which were made about the heuristics which subjects mentioned. They included justifications or explanations for the use of a heuristic, conditions for the applicability of a heuristic, and the power or usefulness of a heuristic.

7. Declarative statements. These were statements which referred to general facts about TEMIM, such as the general format of problems, as well as more specific hints or caveats about solving problems which were not classifiable as heuristics. An example of this latter type might be: "Remember, you don't need to use all of the parts in a problem".

7.3.1 Discussion

The recipe test results in Table 11 proved to be inconclusive in pointing to any clear differences between the subject groups. The variation among subjects in general was such that comparisons among the groups were often not meaningful. These tests, like the experimental task itself, were relatively loosely constrained, and thus allowed for the

expression of the wide variety of results observed. The results presented here do not do justice to the rich variation in the data collected.

The groups' use of the heuristics uncovered in the recipe tests reveal no obvious associations between the various subject groups and any specific heuristics. It is somewhat curious that only one subject explicitly mentioned the subgoaling strategy. The informal

TABLE 11. Use of heuristics, metastrategic statements and declarative statements for the final recipe test (columns 1-5 are yes/no, columns 6 and 7 are counts).

Subject	Spatial	Treatment	1.	2.	3.	4.	5.	6.	7.
SJ	High	1						2	5
CS	High	1			✓		✓	2	0
JB	High	1	✓		✓		✓	1	0
KC	High	1	✓					1	0
JH	High	2		✓				0	1
MT	High	2	✓				✓	0	2
TB	High	2	✓					0	0
FN	High	2						0	4
PK	Low	1						0	8
JG	Low	1	✓	✓				0	5
DG	Low	1						0	0
ML	Low	1				✓	✓	2	0
PP	Low	2		✓				0	2
AS	Low	2		✓	✓		✓	0	2
JO	Low	2						0	1
DN	Low	2	✓				✓	0	2

observations made during the experiment indicated that this was one of the most important strategies used by subjects. The results of the frequency counts of metastrategic and declarative statements are also difficult to interpret and present no clear patterns. One notable exception is that while all of the high spatial Group 1 subjects made at least one statement in the metastrategic category, only one other subject from the remaining three groups made any such statements. Individual differences seem to be the most likely source of these results. In sum, the main contribution of the recipe tests was in revealing and confirming some of the general strategies and heuristics that subjects were observed to use over the course of the experiment.

7.4 INFORMAL OBSERVATIONS

Consistent with the exploratory spirit of this study, this section will briefly describe some of the interesting or unusual phenomena which did not emerge in the discussion of the other results. Most of these results are based on written observations made during the course of the experiment.

One of the most prevalent and interesting phenomena not captured by the analyses presented above was the occurrence of instances of fixation. Fixation was clearly a significant cause of impasses for subjects throughout the study. These episodes tended to occur in one of two forms: subjects would either become fixated with respect to the function of a certain object or objects, or would become fixated with respect to the configuration of a given group of objects. In the first case, subjects would persist in associating an object or objects with certain functions, failing to consider the object(s) in terms of additional affordances which were relevant in the current context. For example, subjects might regard a pair of scissors exclusively as a cutting tool in a case where the problem required them to be used as a barrier of some kind. In the second case, the fixations occurred with respect to the specific arrangement of objects , where subjects would fail to recognize all of the relevant degrees of freedom available in arranging objects. This sort of fixation could be related to a failure to accurately separate the "figure" of a problem (i.e., what parts could be manipulated), from the "ground" (i.e., which parts could not be manipulated). Subjects would become fixated on attempting to modify certain objects in an effort to make a given system operate in the desired manner, while completely ignoring other objects which could in fact be manipulated much more productively with respect to the current goals. Even in the face of explicit attempts to find alternate ways of approaching a problem, subjects would fail to consider the relevant objects.

Two important ways of generating ideas or "seeds" for solutions to problems or subproblems were simulating and dragging objects through the workspace. Occasionally subjects would use the simulation feature for the explicit purpose of trying to generate ideas about how to proceed in a given problem. This would often consist of simply simulating repeatedly and observing the system in motion. Subjects made statements such as "I'm just running the machine to get ideas for what to do next". The second such method was similar in intent except that it focused on a specific object. Subjects would drag one of the user-configurable objects through various positions in the workspace in an attempt to cue ideas for how it might be used, presumably through the act of observing the object in proximity to other objects which could potentially be used together with it. While this second method was not as common as the first, it seems to emphasize the importance for

some subjects of being able to visualize potential configurations of objects. Both of these methods were often used when subjects had reached an impasse and had acknowledged that they were not making progress. Thus, these methods served as ways to generate fresh ideas or to break fixations.

Another strategy which many subjects occasionally followed was to appeal to what might be termed a belief in the "rationality" of the problems, particularly in the case of impasses or situations where subjects were having difficulty deciding what to do next. Because they knew that they were dealing with contrived problems which had solutions (recall that they were told that all problems were solvable), subjects would sometimes regard certain objects, either in the parts bin or among the pre-configured objects in the workspace, with the assumption that "they must be there for a reason". Proceeding from this assumption they would try to construct subsystems involving the objects concerned, perhaps hoping that the results would bring them closer to the goal. The ultimate goal of the problem seemed to have very little direct influence on this type of reasoning. Rather the use of this assumption was a way to try make progress blindly by arranging the given objects in what seemed like a "rational" manner, with the assumption that the designers of the problems were themselves rational and therefore would have constructed the problem in a consistent and comprehensible manner. When searching for a way to approach a problem, whether as a reaction to an impasse or even from the outset, this strategy helped subjects to focus their efforts.

8. GENERAL DISCUSSION

The most salient results revealed in the various analyses were arguably the definitive differences in trial completion time both between the high and low spatial groups and between the baseline and reduced constraint groups. While these differences in performance were relatively clear, the results proved much more difficult to interpret with respect to the factors which may have influenced these differences. In the following sections, the most significant results are revisited and an attempt is made to integrate some of the findings. A speculative model is proposed as a mechanism by which spatial ability and the constraint manipulations may have expressed themselves in the performance differences observed.

8.1 SPATIAL ABILITY AND CONSTRAINT

In examining the results as a whole, spatial ability was clearly an influential factor. At a basic level, problems in TEMIM deal with arranging and orienting objects in space, subject to various forms of constraint. Moreover, solving problems invariably deals with producing appropriate movements of objects within this space to achieve various functional goals. Thus, proper functioning of systems of objects depends intimately on their spatial configuration. Given these premises, it is not difficult to come to the conclusion that spatial ability is likely to be an important factor in subjects' ability solve problems.

However, the precise role that spatial ability played in facilitating performance was difficult to determine from the results. Authors such as Hegarty (1991) and De Kleer and Brown (1983) have stressed the importance of spatial processes in inferring the behaviour of existing mechanical devices. This type of inference can be compared to mentally "simulating" the behaviour of existing configurations of objects in TEMIM, a process which would be expected to produce the same sort of feedback (although less accurately and completely, and potentially involving far greater cognitive effort) as actually using the simulation feature. This was the basis for the assumption that subjects of low spatial ability might be more reliant on the simulation feature to provide this sort of information, resulting in a tendency to use the simulation feature more frequently. The results failed to show that spatial ability and simulation use traded off in this manner. In fact, the rate of simulation seemed to be more closely associated with performance for the high spatial subjects than the low spatial subjects. Moreover, large proportions of subjects' use of the simulation feature were found to be associated with episodes of "fine tuning", which made it less clear

what inferences could be drawn about the relationship between simulation use and spatial ability. In general however, using the simulation feature involved very little cost and often provided considerable informational benefits, making it adaptive for all subjects to use it frequently.

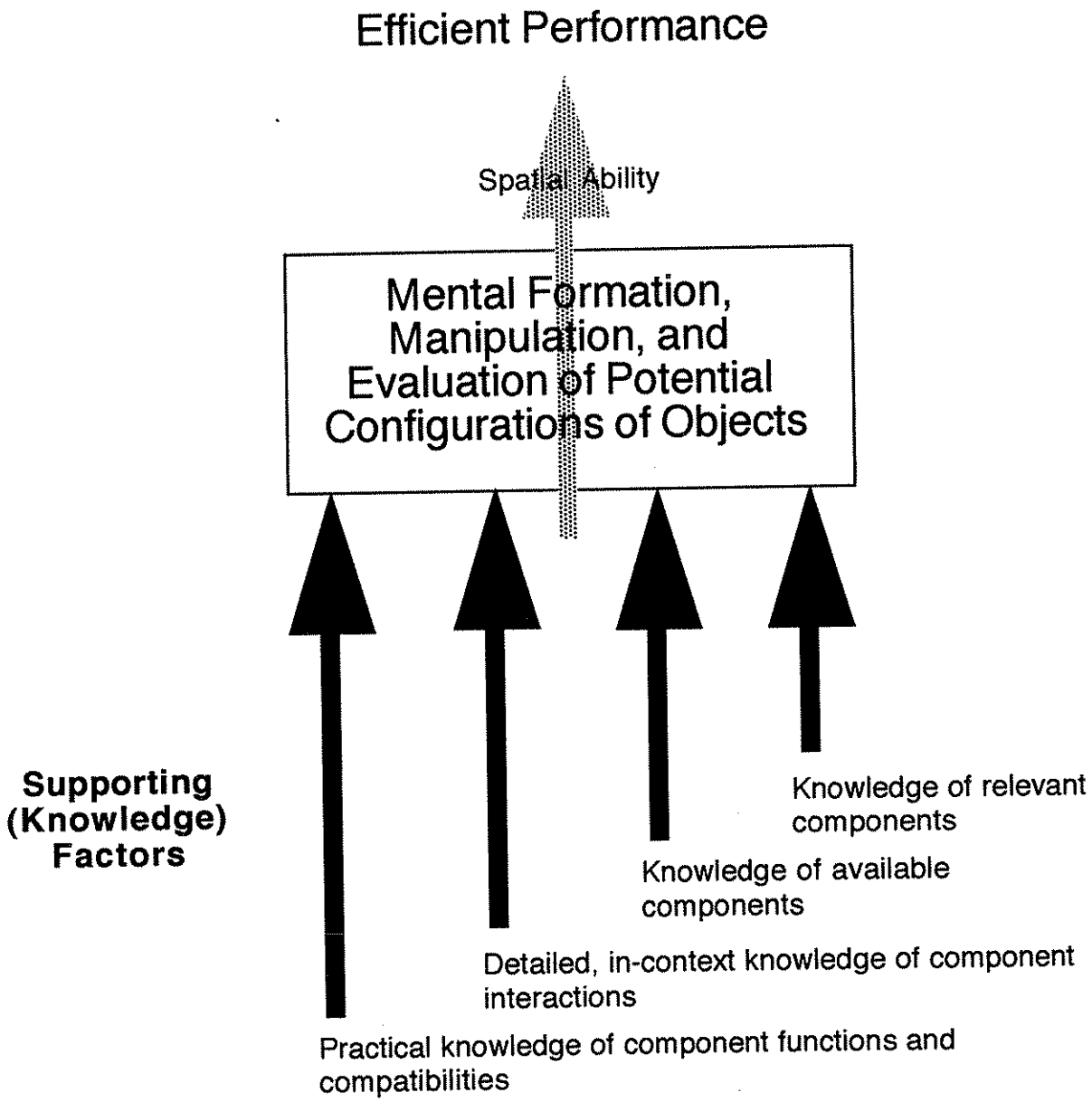
In any case, it seems reasonable to attempt to identify other mechanisms by which spatial ability may have mediated the large performance difference between the low and high spatial groups. One possibility is by facilitating subjects' ability to mentally construct and evaluate future, potential configurations of the available objects in a given problem. Such an ability to "look ahead", quickly mentally constructing and making gross judgments about the feasibility of potential configurations would clearly be useful in helping subjects to make more productive choices about which solution paths to pursue at any given point. Note that this would not necessarily involve forming a representation of an entire solution, nor would it even require motivation by any specific subgoal, rather it would simply involve a more or less distant extrapolation from the current configuration based on a process of spatially representing, manipulating, and potentially animating configurations of objects. This process clearly implicates spatial visualization ability.

It is possible to speculate that this sort of "lookahead" process may in fact have been central in allowing subjects to solve problems efficiently. Note though that while spatial ability would be a key factor in performing lookahead, it is not sufficient by itself to support this kind of process. As Hegarty (1991) has pointed out, the influence of knowledge factors must also be considered in spatial reasoning processes. Accurate causal ("practical") models of how objects function together (i.e., in terms of their affordances and interactions) would be crucial to the effectiveness of such representations by allowing subjects to construct accurate qualitative causal models of subsystems of objects. Any such model however, would be less useful to the extent that there was uncertainty regarding the components available for use in constructing systems. It would also be less useful as a result of uncertainty regarding which of the available set of components were in fact relevant to solving the problem at hand. Even with such knowledge, as De Kleer and Brown (1983) have noted, direct evidence is usually required to disambiguate many details of such models. Given these knowledge components as potentially influential factors in shaping the feasibility and efficacy of lookahead episodes, it becomes possible to propose a simple, speculative[4] model of the lookahead process which, in addition to offering a mechanism for the influence of spatial ability on performance, also offers an account of one

[4]It is important to emphasize the speculative nature of this model. There is no direct support for this model from the experimental evidence. It is offered as one potential integrative framework for interpreting some of the results, which may serve to generate hypotheses for future research.

possible mechanism through which the constraint manipulations may have influenced performance. This model is illustrated in Figure 10.

FIGURE 10. A model of the factors facilitating the lookahead process.



The diagram illustrates the four types of knowledge proposed as supporting factors for the lookahead process, while spatial ability is represented as the primary factor facilitating the actual execution of lookahead. Note that for each of the knowledge factors, it is inappropriate to say whether subjects do or do not have this knowledge; there is no one piece of unambiguous knowledge which corresponds to each factor. There may be "bugs"

or "gaps" or varying levels of uncertainty in subjects' knowledge with respect to all of these factors. The point is that each one supports the efficacious use of lookahead to the extent that it is accurate and complete. In all but the shallowest episodes of lookahead, it seems reasonable to assume that spatial ability would be crucial; the number of changes in interacting spatial (and spatially dependent) constraints associated with mentally adding or moving components within a subsystem of objects would likely be completely intractable for subjects to manipulate in any sort of propositional form.

Interestingly, the constraint manipulations conducted in this experiment, which reduced or hid constraint to some degree, each map onto one or more of the knowledge factors proposed. For CB trials, both knowledge of the available parts, and thereby knowledge of the relevant parts were hidden from subjects in Group 2. This would have severely eroded the basis for subjects' ability to perform lookahead in ways relevant to the problem. Subjects could of course perform limited episodes of speculative lookahead using parts which they imagined might be useful. In fact, this may often have been the basis for subjects' requests for certain parts. The fact that two of the proposed supporting types of knowledge were hidden from subjects during these trials may have contributed to the fact that CB trials were performed most slowly among all trial types.

During RS trials, Group 2 subjects were discouraged, and eventually blocked, from collecting knowledge about the detailed, micro-level interactions between objects. Because the results of lookahead would be highly dependent on the assumptions subjects made regarding the base configuration from which they were extrapolating, this type of knowledge would have a large influence on the accuracy of lookahead episodes. For trials not subject to the restrictions in RS trials, subjects could gain perfect information about the base configuration by simulating, but for cases where they were discouraged or blocked from doing so, the effectiveness of lookahead would clearly be reduced.

During IP trials, the uncertainty about the relevant parts was increased for Group 2 subjects and decreased for Group 1 subjects to the point where almost all of the available parts were actually relevant. Note that regardless of how many parts were in fact relevant in constructing a solution, subjects would necessarily not be certain of which parts were relevant and which were not. However, increasing the number of parts substantially could nonetheless be expected to considerably increase this type of uncertainty given the same pre-configured scenario.

The one knowledge factor which does not have an obvious association with any particular trial type is practical, causal knowledge of the functioning of objects and their interactions. This factor could instead be associated with learning processes and how much subjects knew or were able to remember about exactly how objects functioned and could be

used in combination with other objects. Such processes were not specifically examined in the current study, and thus the impact of this type of knowledge cannot be assessed here. Nonetheless, in the case of practical knowledge and for the three other types discussed, the effectiveness of the extrapolation performed during lookahead would be based on the extent and accuracy of the available information relevant to the current situation[5].

Returning specifically to the results concerning spatial ability, the interaction between spatial ability and the constraint manipulations seemed to be illustrated most clearly in the correlations of spatial ability score with the product and various process measures. These results suggested that in general, reduced constraint conditions effectively eliminated the influence of spatial ability on many of the measures. Recall that while many of the process measures were significantly correlated with spatial score for the baseline group, none of the measures were significantly correlated with spatial ability score for the reduced-constraint group. If the main impact due to spatial ability was associated with the facilitating effect it had on lookahead, then manipulations which made lookahead difficult for other reasons would be expected to reduce the effects of spatial ability on these factors.

As alluded to in previous sections, this sort of process may also have been the cause of the seemingly anomalous result that high rates of simulation were more highly related to good performance for high spatial subjects than low spatial subjects. The ability to apply information regarding constraints gained through simulation to envisioned future scenarios might be expected to be far more influential in shaping performance than the use of simulations merely as a source of feedback about the adequacy of the current configuration[6]. Thus, to the extent that subjects were able to use the information provided by simulation in this way, their performance would be expected to be more closely related to the rate of simulation. Although no data were collected on the previously noted episodes of subjects using the simulation feature explicitly to "generate ideas", it is also possible to predict on the basis of the suggestions above that these episodes may have been more prominent for high spatial subjects.

---

[5]Interestingly, the proposed knowledge components supporting lookahead bear considerable resemblance to basic models of the types of information required to support knowledge-based computer systems designed to perform configuration tasks (Stefik, 1995).

[6]Collectively, when taken in the light of the informal observations made during the experiment, these results suggest two qualitatively different modes of simulation usage: simulation to discover and suggest gross constraints on the feasible configurations of parts, and fine tuning of spatial arrangements of parts in an effort to successfully trigger or continue a desired causal chain. The difference may not lie so much in the intended purpose of a simulation run, but rather in the type of information which subjects attended to and/or spontaneously picked up on during the course of simulation. The analyses performed in this study have confounded these modes, perhaps leading to some of the less intuitive results obtained.

8.2 ABSTRACTION

Turning to the subjects' use of abstract references, the most salient result with respect to this measure was the fact that the rate of use of such statements seemed in general to be correlated with poor performance rather than superior performance. The actual rates of such statements were in fact quite low, although there was extreme variation between subjects. The relative scarcity of such statements may be related to the fact that solving problems in TEMIM is so closely associated with performing operations in the workspace, manipulating the available parts. The available representation of the problems (i.e., the workspace) is inherently concrete, which may have encouraged subjects to form and maintain their own representations of the problems more consistently at this level than for other types of problems which are represented more abstractly.

It is somewhat more difficult to propose a reason for the fact that abstract references seemed to be more frequent for subjects who performed relatively poorly, or for the fact that this effect was much stronger for the reduced-constraint group than for the baseline group. One possibility is that subjects may have used abstract references primarily in response to difficult scenarios.

8.3 GOAL CHAINING AND ASSOCIATIVE STRATEGIES

Goal chaining was in fact not as prevalent as a general strategy as originally expected for some interesting reasons. Consistent episodes of forward or backward chaining in the construction of a causal sequence were in fact relatively rare. Combinations of the two within problems were much more common, along with episodes of purely associative part placement. In fact, subjects seemed for the most part to become reasonably well attuned to efficient ways of reducing the degrees of freedom within a given problem scenario. Various permutations of the strategies mentioned here, in concert with others, possibly including lookahead, were generally used frequently, as the situation and abilities of the subjects dictated.

Associative strategies, which several subjects mentioned in their recipe tests, are meant to describe episodes where subjects chose objects from the parts bin and placed them in the workspace based primarily on the perceived association with the parts already in the workspace. Thus, the choice of an object to be added to the current configuration at a given point was driven only by knowledge of which objects were typically associated with those already in place. For example, given a magnifying glass somewhere in the

workspace, many subjects would immediately look for a flashlight, simply because that was the object most commonly associated with magnifying glasses. This was often done with no apparent consideration of the purpose which the magnifying glass / flashlight combination would ultimately serve.

It is not unreasonable to assume that the cognitive effort involved in performing the "lookahead" process described earlier is considerably more than for simple forward or even backward chaining. In particular, the effort involved in lookahead would be expected to grow exponentially as the qualitative "distance" between the current and envisioned configurations grew[7]. If one assumes that this is so, it is not difficult to see the relative cognitive economy, even over the chaining methods, that is offered by the use of associative methods. Moreover, in many problems in TEMIM, an associative strategy can in fact serve to considerably reduce the initial degrees of freedom in a new problem. Because the number of available objects in any given problem is typically small (on the order of 10 objects), it was often possible for subjects to choose an initial placement for a small number of parts based purely on associative grounds (i.e., with no consideration of purpose), and thereby greatly aid in constraining the form of the remaining portions of the solution.

Sometimes, large portions of problems could be solved at once in this way. Certain arrangements of objects seemed for some subjects to be the basis for identification of "prototypical" solutions to common subproblems. Subjects seemed occasionally to "recognize" certain situations and directly associate them with certain configurations of parts as solutions. In some sense, this can be compared to the recognition-based problem solving techniques described by Klein (1989).

These sorts of associative processes were equally relevant when subjects reached an impasse. Resorting to an associative strategy could often provide the small amount of constraint necessary for subjects to be able to continue making progress towards a solution. It may be that these associative methods were an effective way of reducing the causal "gaps" in the system, perhaps to the point where processes such as lookahead could be employed.

Of course, excessive reliance on an associative chaining to determine object selection and placement made subjects highly susceptible to blindly following "dead end" solution concepts. Very few problems were so highly constrained that a purely associative strategy would in fact lead to the goal without considerable need for backtracking and re-

---

[7]No concrete definition of this "distance" is proposed here, although one could easily imagine such a definition based on differences between the current and envisioned configurations in terms of the number of parts used, differences in placement and orientation of parts used, etc.

evaluation. Therefore, although most subjects did at times use this strategy to their advantage, it was not sufficient by itself to allow subjects to achieve good performance. Better subjects seemed to be judicious in their use of associative chaining.

Subjects' tendency to appeal to the "rationality" of problems (discussed in section 7.4), can be viewed in a similar light. Similar to associative chaining, this strategy was one for which cognitive effort seemed to be relatively low compared with other potential methods of trying to make progress in situations where then next action to take is unclear. This method would have relieved subjects of trying to perform any significant amounts of lookahead, and in fact seemed often to be associated with situations where looking ahead failed to suggest productive actions to take or where lookahead may have been especially difficult. Of course this strategy was also prone to leading subjects towards "dead ends" when the objects subjects chose to attempt to include under this assumption were in fact distractor objects offering no progress towards the goal. Nonetheless, it was frequently also useful as a way of generating ideas through the process of considering what could be done with the objects identified.

With respect to the constraint manipulations, the most salient result regarding goal chaining seemed to be the fact that the rate of goal definitions was much more highly correlated with spatial ability for Group 1 subjects than for Group 2 subjects. If the assumption that the primary role of spatial ability is in facilitating lookahead is maintained, then this result would be expected since the knowledge factors which allow for the expression of spatial ability during episodes of lookahead were eroded for the majority of Group 2's trials.

9. LIMITATIONS

Within the scope of exploring the use of a novel task as a domain for problem solving and design research, this study yielded a number of interesting results and raised many questions for continued research using this task. However, the current study was limited in several important ways which should be considered in evaluations of the results and in the planning of future work.

A significant shortcoming of the current study was the lack of any validated means of classifying problems. The problems used in the current study were chosen somewhat arbitrarily, although based partially on the results of the pilot study and other previous experience with TEMIM. The process of constructing such a taxonomy would likely prove to be informative in itself, but would also enable more informed choices of problems to include in future studies[8]. Being able to classify problems in a meaningful way would perhaps allow further relations to be drawn to previous work, and provide a more objective basis for choosing problems to present to subjects in future studies. It would also allow for meaningful evaluations of the nature of the relationship between problem difficulty and performance, and could eventually lead to a cognitive task analysis of TEMIM, or individual problems in TEMIM.

The partially ill-structured nature of TEMIM is, in one sense, one of its strengths. This characteristic is one of the factors which separates it from traditional laboratory research on problem solving, and allows for the expression of a wide variation in, and unique examples of, subjects' problem solving techniques. However, because solutions to problems are typically non-unique, this makes it difficult to comprehensively evaluate subjects' performance, and may have served to confound some of the measures in the current study. No means were developed in the current study, other than completion time, for classifying the quality of subjects' solutions. While the number of parts used was suggested as a possibility, this remains a superficial assessment of the quality, or elegance, or simplicity of subjects' solutions.

One of the primary limitations of this study was the small sample size. With only 16 subjects in total, it is impossible to discount the fact that some of the results may have been spurious or due to coincidental individual differences. There is a significant tradeoff to be considered in any reasonably complex domain between investing the time to train (in a controlled way) a sufficient number of subjects until their performance has stabilized, and

[8]It should be remembered that another of the great strengths of TEMIM is that it is possible to modify the supplied problems or to create entirely new problems.

attempting to locate subjects who may already have experience (of a type and amount you cannot closely control) within the task domain.

In a related point, many of the results in this study were potentially confounded with learning phenomena. It is possible, or even probable, that Group 2 subjects were unprepared for the various constraint manipulations under which they were asked to perform, and spent significant amounts of time attempting to adapt their strategies. It was also apparent among many subjects that their causal models of object behaviours and affordances were incorrect or incomplete, even deep into the experiment.

Another limitation in many of the analyses was that all of the constraint manipulations were treated as equivalent. Although some of the results made reference to the effects of individual types of manipulations, the differences between them were not explored explicitly. Although the lookahead model proposed previously suggests ways of exploring the differences between the various trial types, the analyses in the current study did not, for the most part, attempt to distinguish among them.

10. ISSUES FOR FUTURE WORK

The main objective of the current study was to raise questions, both with respect to design problem solving and problem solving in general, which could conceivably be pursued in future research using TEMIM. This section will briefly touch on a number of these as well as some possible techniques for inquiring about some of the issues raised.

The protocols collected during the introductory sessions were not analyzed at all, and may be a fruitful source of data on how subjects became attuned (or mis-attuned) to the constraints of the TEMIM environment. As one example, it would be interesting to track the development of subjects' use of common subsystems of objects as solutions to common subproblems. It is possible that subjects differed widely in their tendency to form and use functional "units" of groups of several objects. Another example of a learning-related issue which could be explored is the development of subjects' strategies over time. Researchers such as Patel, Groen and Arocha (1990) have suggested that subjects' reasoning tendencies shift from backward chaining to recognitional, associative, forward chaining strategies as expertise develops. It might also be interesting to examine the development within subjects of the perceived salience of objects' affordances as expertise develops, both in terms of the order and number of affordances subjects associate with objects.

Another potential technique for examining learning within subjects, which could also be used to explore expertise differences between subjects, is a "categorization test" similar to that used by Chi, Feltovich, and Glaser (1981). It might prove informative to show subjects initial problem scenarios and have them classify these based on their initial interpretations of how they might be solved. Although evaluation may prove to be difficult due to the often wide flexibility in solution methods, such a test may reveal considerable differences in how subjects form their initial problem representations, both across time and across subjects.

Given a cognitive task analysis of a set of problems in TEMIM, future studies could be structured specifically to examine strategies and transitions between strategies. An improved means of examining subjects' strategies could be used to better evaluate the precise effects of constraint manipulations such as the ones performed in the current study. It might be valuable to construct abstraction hierarchies for individual problems and attempt to map subjects' reasoning onto such representations (see, e.g., Vicente et al., 1995). However, the results of the current study have raised doubts about the general coherence of subjects reasoning episodes, pointing more towards the use of less structured,

opportunistic tendencies, as well as the relative lack of any extensive, overt use of higher levels of abstraction.

The protocol analysis performed in the current study was selective, classifying only a relatively small proportion of subjects' verbal activity. It may be valuable in the future to fully transcribe one or more trials along with detailed behavioural protocols for all of the subjects from the available data (or from subjects in a future study) to attempt to examine more fully the nature of subjects' reasoning.

After each trial in the present study, subjects were asked to briefly describe how they solved the problem. These post-hoc explanations were not analyzed but may be a very rich source of data. They were difficult to code in the current study because subjects varied considerably in the amount of information they were able to provide about how they had solved problems. However, despite this and other limitations of such retrospective reports, such analyses may reveal much about how deeply subjects' understood the constraints operating in a given scenario, how well they understood their own strategies, and may provide clearer pictures of how subjects assessed and approached problems and dealt with impasses (see the results on "knowledge encapsulation" in Boshuizen and Schimdt, 1992). As an example of the sort of information such an analysis could provide, it was noted during the experiment that subject CS of the high spatial Group 1 subjects made considerable use of abstract references during these post-trial explanations. Although CS also used relatively high rates of abstract references during trials, his use of such statements was clearly much more frequent during the explanations. This may indicate that he was in fact more frequently representing problems in terms of higher levels of abstraction than was revealed by the concurrent protocols.

Also collected but not analyzed in the current study were semi-structured post-experiment debriefings where subjects were asked about various aspects of their strategies. Some of the discussions were quite extensive and provided many interesting insights. The results from these could prove quite illuminating, and could potentially be compared against and treated as elaborations of subjects' recipe test results.

In the discussion of how well TEMIM problems represent design tasks, it was noted that although subjects were not given any explicitly conflicting criteria for performance in the current study, they may have spontaneously adopted some. It would be possible to make such requirements explicit, for example by telling subjects that they would be evaluated based on not only the time taken to solve problems, but also the number of parts used in solutions, the number of uses of the simulation feature, or the number of distinct solutions attempted (or the amount of backtracking within solutions). Giving subjects each of these as explicit performance criteria would create a situation involving

multiple, conflicting constraints to be satisfied at once, a situation much more typical of most design problems.

One of the interesting phenomena observed in the current study was the tendency for subjects to attempt to decompose problems into smaller subproblems. Simon (1973) has noted this as a prototypical method of dealing with complex, ill-structured problems such as those typically encountered in design. It might be very interesting to perform a study specifically designed to elicit and examine instances of problem decomposition. How subjects perform such decompositions, on what bases, and the interaction of such episodes with other problem solving strategies could all prove informative. Differences might be found in how well subjects are able to decompose problems into independent subproblems and what sorts of phenomena arise when subjects are not able to do so. For instance, subjects in the current study were often observed to make errors due to treating interacting parts of problems as though they were independent. They would make changes to one portion of a problem only to find that the changes affected another part of the problem in undesirable ways. The relative ability of subjects to perform such decompositions in ways which facilitate effective problem solving might be shown to be a significant determinant of performance.

The present study did not extensively explore the precise nature of the effects that the various constraint manipulations had on subjects' strategies. It may prove informative in the future to try to examine in more detail exactly what the strategies used by subjects under each trial type were and how these were related to the constraint manipulations.

The previous discussion of fixation phenomena failed to touch upon a related phenomenon which, although difficult to identify with any certainty, seemed to be relatively frequent, that is, insight (or "restructuring"). There is some debate about precisely what this term means and, in some instances, even about whether or not it exists (see, e.g., Weisberg, 1995). Duncker (1945), characterized insight as reformulating or redefining the goal and/or the givens in a more "productive" way (i.e., one that makes it more apparent how the givens can be used to achieve the goal). There is general agreement about the observable characteristics of what is normally referred to as insight, such as a sudden "Aha!" type of statement (Hadamard, 1945; Davidson, 1995; Kaplan and Simon, 1990) followed by either a new goal definition or a directed sequence of problem solving activity. It is obviously very difficult to define a rigorous test of whether or not an instance of insight has actually occurred, but these observable behaviours can be potentially used as an operational definition.

It is likely that different forms of restructuring may be observed corresponding to the forms of fixation noted in the current study. Both fixations involving object

affordances, and fixations related to the subjects' parsing of problems into figure and ground may give rise to insight phenomena. Ohlsson (1984) has suggested that impasses such as those caused by fixations may cause subjects to engage in an explicit search for alternate ways to restructure their problem representation. This sort of hypothesis could potentially be tested in the TEMIM environment. It might also be interesting to examine how restructuring events associated with object affordances serve to influence the relative salience of those affordances for subjects. It may also be possible to test ideas about subjects' tendencies to perceive objects directly in terms of their affordances. When objects become synonymous with their affordances, it may be that restructuring is required to break the association; that is, "direct perception" of some affordances in favour of others may actually hinder performance if the affordances perceived in a given situation do not include the entire relevant affordance set. It may be possible to observe the process of subjects becoming attuned to the constraints represented by the affordance set for particular objects, which may in turn get tuned in the context of restructuring events.

Because the nature of problems such as those in TEMIM often requires novel uses of objects and systems of objects, insight-related phenomena such as functional fixedness, Einstellung, incubation, and restructuring may all be fundamental in problem solving of this kind. Thus, processes of problem restructuring and re-representation, such as those first discussed by Duncker (1945) and by Selz (in Frijda and De Groot, 1982), might prove to be very valuable templates on which to map such behaviours in TEMIM. The recent resurgence in interest in insight phenomena (see Sternberg and Davidson, 1995) has led to a simultaneous resurgence interest in the work of these and other of the Gestalt psychologists, such as Wertheimer, Kohler, Luchins, and others. The work of these researchers is perhaps particularly relevant because in general, they chose to study problems of, it can be claimed, practical significance, or at least problems more complex than much of the laboratory research on problem solving done since. Newell and Simon (1972) explicitly stated their lack of consideration of the types of phenomena observed by the Gestaltists.

Despite the criticisms leveled at the Gestalt movement regarding their lack of direction and rigorous methodology, there is reason to reconsider their findings as interest turns to more complex forms of problem solving, such as was the intent in the current study. A recent quote from Mayer (1995), illustrates the sentiment well: "Sometimes the Gestaltists are accused of being soft scientists whereas contemporary psychologists consider themselves workers in a hard science. An alternative way to frame the distinction is that the Gestalt psychologists work on hard questions, whereas modern cognitive psychologists sometimes prefer easy ones" (p. 26).

Lastly, the "lookahead" model introduced in section 8.1 remains as a largely speculative proposal, with only limited support from the data collected. In order to gain any confidence in this model, it needs to be tested directly. Variations on the constraint manipulations performed here could be used to perhaps more directly test each of the knowledge factors proposed. It might also be possible to devise tests to determine whether subjects are explicitly engaging in lookahead at all, and if so, how far ahead they are looking, what factors are guiding such episodes, what factors triggers such episodes, how subjects act based on the results, and so on.

The apparent similarity of the model to some basic accounts of knowledge-based configuration systems (Stefik, 1995) may be worth pursuing. While the similarity of how subjects actually execute lookahead to how such AI-based systems perform configuration tasks (i.e., search) may be limited, there may nonetheless be many other similarities and insights to be gained from a more thorough comparison of the two, especially as to how knowledge affects the ability to perform such processes. Such a comparison may make it possible to form more specific hypotheses or to more clearly interpret some of the results presented here.

11. CONCLUSIONS

This study presented a novel task domain, TEMIM, which was proposed as a design "microworld". The study undertook to explore the nature of subjects' problem solving behaviour in this type of domain with the ultimate goal of providing direction for aiding practitioners confronted by similar types of problems, primarily through suggesting potential avenues for continued research using this or similar tasks.

Several similarities were noted between the task characteristics and characteristics identified as typical of design problems. In particular, many aspects of problems in TEMIM are relatively ill-defined, especially in comparison with previous laboratory research on problem solving. On the other hand, few previous studies of design behaviour have attempted to isolate design problem solving at the level of the individual and to observe such behaviours under relatively structured laboratory conditions. In allowing this, TEMIM offers a potentially fruitful domain for studying subjects' performance on design-like problems.

The results revealed that spatial ability is a highly influential factor shaping subjects' performance. Also found to be influential were three distinct types of manipulations which either reduced or hid constraints from subjects in some way, effectively increasing the extent to which the problems were ill-defined. A speculative model describing a "lookahead" process was proposed as one possible mechanism by which these factors may have combined to produce the results obtained. The model specifies certain types of knowledge which must be present in more or less certain form to support the lookahead process, which operates on spatial representations of envisioned configurations of objects. While the model is consistent with the data, it is speculative and needs to be tested directly in further work. Also, the model clearly does not account for all relevant aspects of subjects' behaviour, nor is it intended as a comprehensive explanation of all of the results. Nonetheless, this model seems as though it could provide a basis for further explorations using TEMIM.

Goal-directed reasoning, while an important technique, was found to be complemented by a variety of other strategies and heuristics such as associative matching and actions based on the assumed "rationality" of the problems. In general, subjects seemed to become well-attuned to efficient ways of reducing the degrees of freedom and uncertainty associated with any given problem state, based on their own preferences and abilities. This meant that small episodes of many strategies would often occur within the same problem as the solution developed and new information was discovered.

The broad, exploratory intent of this study and the fact that no specific hypotheses related to design were tested means that any attempt to draw specific conclusions which can be directly applied to design would be tenuous and premature. The main contribution of the study is in uncovering some of the general characteristics of subjects' performance in this domain. This, it is hoped, will provide the basis for continued efforts to better understand the nature of design problem solving and, more importantly from a cognitive engineering standpoint, how to better support designers in their tasks. TEMIM certainly represents only a small section of the conceivable space of "design" problems, and is limited in significant ways. However, the type of problems and problem solving elicited here will, it is hoped, prove to be a step towards the systematic study of this and other important forms of design.

REFERENCES

Akin, O. (1984). An exploration of the design process. In N. Cross (Ed.), Developments in design methodology, (pp. 209-220). Chichester: Wiley.

Boshuizen, H. P. A., and Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. Cognitive Science, 16, 153-184.

Burns, C. M. and Vicente, K. J. (1994). Human factors design guidance: Matching the advice to designers' questions (Contract report XSE93-00010-(303)). Toronto, Canada: Cognitive Engineering Laboratory.

Carroll, J. M., Thomas, J. C., and Malhotra, A. (1979). Clinical-experimental analysis of design problem solving. Design Studies, (1), 2, 84-92.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Chi, M. T. H., Glaser, R., and Farr, M. J. (Eds.) (1988). The nature of expertise. New Jersey: Lawrence Erlbaum Associates.

Darke, J. (1984). The primary generator and the design process. In N. Cross (Ed.), Developments in design methodology, (pp. 209-220). Chichester: Wiley.

Davidson, J. E. (1995). The suddenness of insight. In R. J. Sternberg and J. E. Davidson (Eds.), The nature of insight. Cambridge, MA; MIT Press.

De Kleer, J. and Brown, J. S. (1983). Assumptions and ambiguities in mechanistic mental models. In D. Gentner and A. L. Stevens (Eds.), Mental models, (pp. 155-190). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Duncker, K. (1945). On problem solving. Psychological Monographs, 58 (5, Whole No. 270).

Dynamix Incorporated. (1993). The (even more) Incredible Machine. Coarsegold, CA: Author.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: ETS.

Ericsson, K. A. and Simon, H. A. (1993). Protocol analysis: Verbal reports as data (2nd ed.). Cambridge, MA: MIT Press.

Frensch, P. A., and Funke, J. (1995). Complex problem solving: The European perspective. Hillsdale, NJ: Lawrence Erlbaum Associates.

Frijda, N. H. and De Groot, A. D. (Eds.). (1982). Otto Selz: His contribution to psychology. The Hague: Mouton.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton-Mifflin.

Goel, V. and Pirolli, P. (1992). The structure of design problem spaces. Cognitive Science, 16, 395-429.

Hadamard, J. (1945). The psychology of invention in the mathematical field. New York, NY: Dover.

Harrison, B., Owen, R., and Baecker, R. (1994). Timelines: An interactive system for the collection and visualization of temporal data. In Proceedings of Graphics Interface'94. Banff, Alberta: Canadian Information Processing Society, 171-178.

Hegarty, M. (1991). Knowledge and processes in mechanical problem solving. In R. J. Sternberg and P. A. Frensch (Eds.), Complex problem solving: Principles and mechanisms, (pp. 253-285). New Jersey: Lawrence Erlbaum Associates.

Hunt, E. (1991). Some comments on the study of complexity. In R. J. Sternberg and P. A. Frensch (Eds.), Complex problem solving: Principles and mechanisms, (pp. 383-395). New Jersey: Lawrence Erlbaum Associates.

Irmer, C., and Reason, J. T. (1991). Early learning in simulated forest fire-fighting. In Simulations, Evaluations, and Models: Proceedings of the Fourth MOHAWC Workshop (pp. 1-20). Roskilde, Denmark: Risø National Laboratory.

Kaplan, G. A. and Simon, H. A. (1990). In search of insight. Cognitive Psychology, 22, 374-419.

Klein, G. A. (1989). Recognition primed decisions. In W. B. Rouse (Ed.), Advances in man-machine systems research (vol. 5). Greenwich, CT: JAI Press.

Lawson, B. R. (1984). Cognitive strategies in architectural design. In N. Cross (Ed.), Developments in design methodology, (pp. 209-220). Chichester: Wiley.

Logan, B. and Smithers, T. (1993). Creativity and design as exploration. In J. S. Gero and M. L. Maher (Eds.), Modeling creativity and knowledge-based creative design. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Malhotra, A., Thomas, J. C., Carroll, J. M., and Miller, L. A. (1980). Cognitive processes in design. International Journal of Man-Machine Studies, 12, 119-140.

Mayer, R. E. (1995). The search for insight: Grappling with Gestalt psychology's unanswered questions. In R. J. Sternberg and J. E. Davidson (Eds.), The nature of insight. Cambridge, MA; MIT Press.

Newell, A., and Simon, H. A. (1972). Human problem solving. New Jersey: Prentice-Hall.

Ohlsson, S. (1984). Restructuring revisited II: An information processing theory of restructuring and insight. Scandinavian Journal of Psychology, 25, 117-129.

Patel, V. L., Groen, G. J., Arocha, J. F. (1990). Medical expertise as a function of task difficulty. Memory and Cognition (18), 4, 394-406.

Rasmussen, J. (1987). Use of computer games to test experimental techniques and cognitive models. In J. Rasmussen and P. Zunde (Eds.), EFISS III (pp. 187-195). New York: Plenum.

Rasmussen, J. (1986). Information processing and human-machine interaction. New York: North-Holland.

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. IEEE Transactions on Systems, Man, and Cybernetics, SMC-15, 234-243.

Rittel, H., and Webber, M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4, 155-169.

Sacerdoti, E. D. (1974). Planning in a hierarchy of abstraction spaces. Artificial Intelligence, 5, 115-135.

Schon, D. A. (1988). Designing: Rules, types, and worlds. Design Studies (9), 3, 181-190.

Siegel, S., and Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York , NY: McGraw-Hill.

Simon, H. A. (1969). The sciences of the artificial. Cambridge, MA: MIT Press.

Simon, H. A. (1973). The structure of ill-structured problems. Artificial Intelligence, 4, 181-201.

Stefik, M. (1995). Introduction to knowledge systems. San Francisco, CA: Morgan Kaufmann.

Sternberg, R. J. and Davidson, J. E. (Eds.) (1995). The nature of insight. Cambridge, MA: MIT Press.

Thomas, J. C., and Carroll, J. M. (1984). The psychological study of design. In N. Cross (Ed.), Developments in design methodology, (pp. 221-235). Chichester: Wiley.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), Foundations of cognitive science, (pp. 527-579). Cambridge, MA: MIT Press.

Vicente, K.J., Hayes B. C., and Williges, R. C., (1987). Assaying and isolating individual differences in searching a hierarchical file system. Human Factors, 29, 349-359.

Vicente, K. J., and Williges, R. C., (1988). Accommodating individual differences in searching a hierarchical file system. International Journal of Man-Machine Studies, 29, 647-668.

Vicente, K. J. and Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics, 22, 589-606.

Vicente, K. J., Burns, C. M., and Pawlak, W. S. (1993). Egg-sucking, mousetraps, and the Tower of Babel: Making human factors guidance more accessible to designers. (Technical Report CEL 93-01). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Vicente, K. J., Christoffersen, K., and Pereklita, A. (1995). Supporting operator problem solving through ecological interface design. IEEE Transactions on Systems, Man, and Cybernetics, SMC-25, 529-545.

Voss, J. F., and Post, T. A. (1988). On the solving of ill structured problems. In M. T. H. Chi, R. Glaser, and M. J. Farr (Eds.), The nature of expertise, (pp. 261-285). Hillsdale, NJ: Lawrence Erlbaum Associates.

Voss, J. F., Wolfe, C. R., Lawrence, J. A., and Engle, R. A. (1991). From representation to decision: An analysis of problem solving in international relations. In R. J. Sternberg and P. A. Frensch (Eds.), Complex problem solving: Principles and mechanisms, (pp. 119-158). New Jersey: Lawrence Erlbaum Associates.

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg and J. E. Davidson (Eds.), The nature of insight. Cambridge, MA; MIT Press.

Wertheimer, M. (1985). A Gestalt perspective on computer simulations of cognitive processes. Computers in Human Behaviour, 1, 19-33.

Woods, D. D., and Roth, E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), Handbook of human-computer interaction, (pp. 3-43). New York: North-Holland.

APPENDICES

APPENDIX A - Introductory Materials for Pilot and Main Studies

Introduction to the Experiment (Pilot Study)

This experiment is part of a study examining subjects' reasoning processes during problem solving activities. As a subject in this experiment, you will be asked to operate a computer game consisting of a series of puzzles. The experiment will consist of 2 sessions. During the first session you will be given an introduction to the game and be asked to complete a set of 21 practice problems. The remainder of the experiment will consist of 11 further problems. You will also be asked to perform a simple off-line pencil and paper test at some point. While operating the game you will be asked to wear a microphone and to attempt to "think aloud" as you solve the puzzles. All trials will be videotaped.

Introduction to the Experiment (Main Study)

This experiment is part of a study examining subjects' reasoning processes during problem solving activities. As a subject in this experiment, you will be asked to operate a computer game consisting of a series of puzzles. The experiment will consist of 5 sessions. During the first session you will be given an introduction to the game and be asked to complete a set of 21 practice problems. The remainder of the experiment will consist of 16 further problems. You will also be asked to perform a series of simple pencil and paper tests at various points. While operating the game you will be asked to wear a microphone and to attempt to "think aloud" as you solve the puzzles. All trials will be videotaped.

CONSENT TO TAKE PART IN RESEARCH

I hereby agree to act as a subject in an experiment entitled:

Research on Human Problem Solving

I have been given a full description of what I shall be required to do in this investigation, and I am aware that I may withdraw from the investigation at any time, and that I have the right to ask in that case for any data collected about my performance, including videotapes of verbal protocol trials to be given to me or destroyed.

I consent to take part in this investigation voluntarily and without any coercion.

Name:_____

Signature:_____

Address:_____

_____

Date:_____

Experimenter: Klaus Christoffersen

Faculty Supervisor: Kim J. Vicente (978-7399)

Demographic Questionnaire

Initials: ____

Sex: ____

Age: ____

Post-Secondary Education (please list current or most recent program first):

Institution	Program	Dates	Specific area of study

Employment (if applicable - please provide a brief description of your current duties):

_____

_____

_____

_____

Have you ever seen either of the computer games called "The Incredible Machine" or "The Even More Incredible Machine" before? (Yes or No) ____

Have you ever used "The Incredible Machine" or "The Even More Incredible Machine" before? (Yes or No) ____

Instructions

Each trial in this experiment will consist of a puzzle which you are to attempt to solve. (All puzzles are solvable). In each trial you will be given a goal to achieve, a pre-existing "scenario" consisting of a number of objects configured in a given way, and a set of objects which you may add to the scenario and configure as you please (within the constraints of the game) to help you achieve the goal. Data will be collected about how long you take to solve each puzzle, how many objects you use in your solutions, and various other measures, but the primary objective in each trial is simply to solve the puzzle presented.

The experimenter will now demonstrate how the game operates. You may ask questions at any time about how to operate the game but the experimenter will not answer questions regarding specific aspects of puzzles or solutions. It is up to you to discover how the objects in each scenario work and interact.

Each puzzle begins with a "puzzle preview" screen which shows the initial scenario and gives the goal which is to be achieved. When you're sure you understand the goal, click anywhere in the preview window to begin.

The screen consists of the play area containing the initial scenario, as well as the "parts bin", which contains the objects which you have available to add to the scenario to try to achieve the goal. The number under each part in the parts bin represents how many individual parts of that type are available. Note that there may occasionally be more parts available than can be displayed in the parts bin at one time. The arrows above the parts bin allow you to scroll through all of the available parts. Note also that you are not required to use up all of the parts given to you.

To add parts to the scenario click on the desired part and drag it to the desired spot in the scenario and click again to release the object. If a red "X" appears over the object, the object may not be placed in that location. Belts and ropes are used in a slightly different manner. To attach a belt between objects, first choose the belt from the parts bin and then click once on the first object to which the belt is to be attached. A red line will then appear emanating from the first object to the tip of the cursor. Move the cursor to the second object to which the belt is to be attached. When the red line turns green, click to secure the belt. If the line remains red, the belt cannot be attached, usually because the parts are too far apart from each other. Ropes are connected in the same way as belts but are different

in that they have unlimited length. Ropes may be run through pulleys by clicking on the pulley before attaching the rope to the second of the two objects. Only one rope or belt may be attached at any one object with the exception of teeter totters, to which ropes may be attached at either end.

Some parts, such as electric fans, need to be plugged into an electrical outlet to operate. When the part is properly plugged in, a plug will appear over one of the outlets. If the part is not properly plugged in, the outlets will appear empty and the part will not work when the machine is run.

Most parts can be modified in one or more ways. Once a part is in the play area, move the cursor over the object to reveal the modifier icons. The garbage can icon at the upper left corner appears with all objects. Clicking on this will send the object back to the parts bin. Red circular double arrows beneath an object indicate that the object may be "flipped", so that it faces in the opposite direction. Simply click on the arrows to flip an object. Double headed blue arrows at the side of an object indicate that the object may be lengthened or shortened to some degree. To do so, click on the arrows and drag the end of the part to the desired spot.

To start a machine, click on the runner in the starting blocks at the top of the parts bin. To stop and reset a machine simply click anywhere. You may start and stop your machine as many times as you wish during each trial.

Verbal Protocol Instructions

Throughout the experiment, you will be asked to try to "think aloud" as you work on each problem. Each trial will be videotaped and you will be asked to wear a headset microphone to speak into. Note that you yourself will not appear on the video recordings. Only the computer monitor and your voice will be recorded. All video recordings will be viewed only by the experimenter and will be kept strictly confidential.

Please try to verbalize as much of your thought process as possible. This is a very important source of data in this experiment, so please make every effort to fully verbalize your thoughts. Thinking aloud might seem slightly awkward and unnatural at first, so the experimenter may occasionally prompt you by saying "keep talking".

If you have any questions about these procedures, please feel free to ask the experimenter.

APPENDIX B - Data Collection Problems

This appendix contains snapshots of the initial configurations of all problems used during the data collection sessions of the main study. Table B-1 contains a list of the goals provided for each problem (as they appeared to subjects), and of the parts included in the parts bin with each problem. Note that for IP trials there are two lists of parts corresponding to the conditions under which each of the constraint groups performed.

TABLE B-1. Goals and Available Parts for each Data Collection Trial

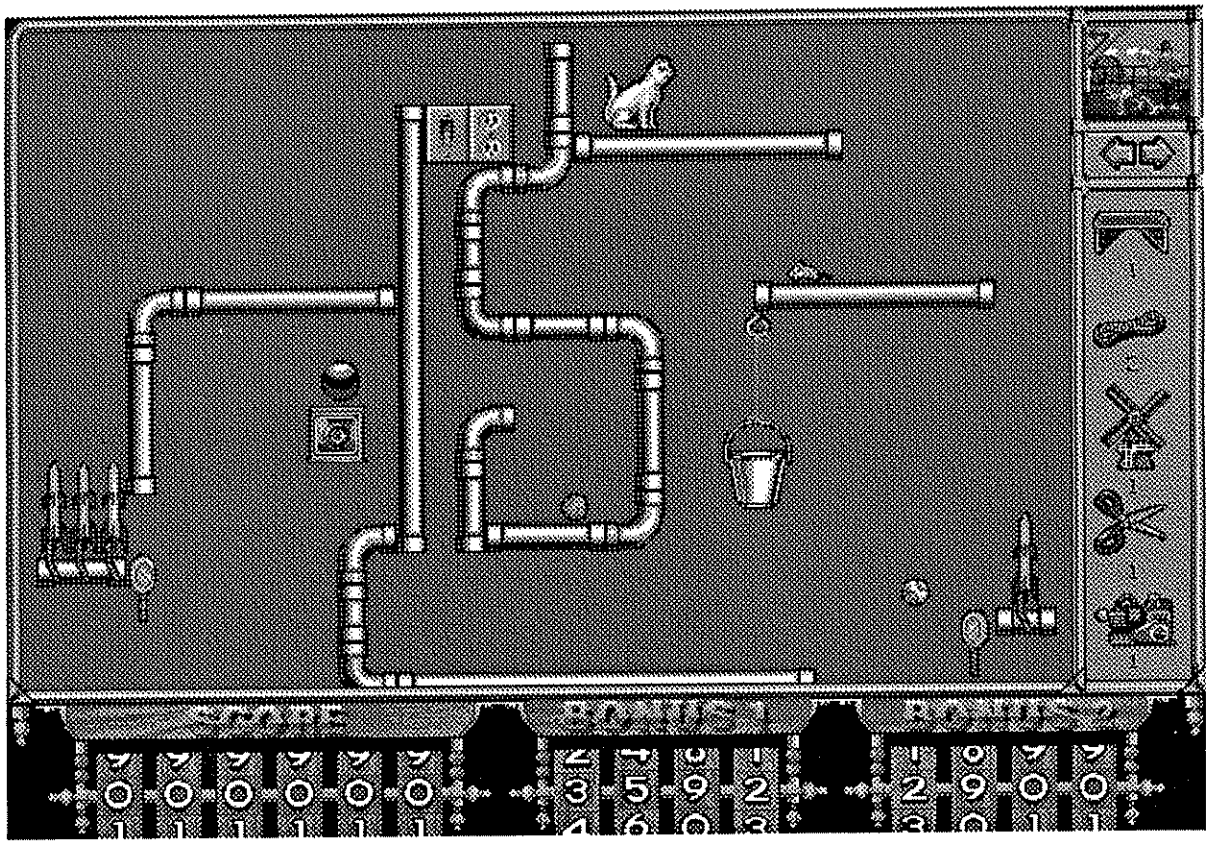
Trial (Type)	TEMIM Problem No.	Goal	Available Parts
1 (N)	54	"Launch all four rockets. All of the rockets must fly off the top of the screen."	1 trampoline 2 belts 1 windmill 1 scissors 1 generator 1 fan 1 electric motor 2 flashlights 1 dynamite/plunger 1 fishbowl
2 (RS)	52	"Your goal is to drop the cage onto Mort the mouse."	1 teeter-totter 1 boxing glove 1 trampoline 1 scissors 1 switch w/outlet 1 fan
3 (IP)	91	"Put the tennis ball in the far left bucket."	<u>Reduced Constraint:</u> 1 bowling ball 1 cannon ball 6 belts 4 gears 1 conveyor 1 windmill 1 rope 3 pulleys 1 generator 1 solar panel 1 magnifying glass 1 flashlight 1 candle 1 mousecage <u>Baseline Constraint:</u> 1 bowling ball 1 cannon ball 2 belts 2 gears 1 conveyor 1 mousecage
4 (RS)	34	"You must get the jack-in-the-box to spring up."	2 teeter-totters 4 belts 6 gears 2 ropes 1 solar panel 1 fan 1 lightbulb

5 (RS)	65	"Make Bob's fishbowl break any way you can."	3 belts 1 windmill 2 fans 1 electric motor 1 magnifying glass 1 dynamite 1 mousecage
6 (CB)	80	"Your goal is to make the basketball go through the hoop."	1 baseball 2 belts 1 jack-in-the-box 1 rope 1 magnifying glass 1 flashlight 1 dynamite 1 monkey bicycle
7 (N)	93	"Punch Pokey Cat."	1 teeter-totter 1 boxing glove 1 belt 1 jack-in-the-box 3 ropes 1 eye-hook 2 pulleys 1 gun 1 scissors
8 (CB)	72	"Put the baseball in the bucket."	1 teeter-totter 3 belts 2 gears 1 rope 1 solar panel 1 fan 1 lightbulb 1 mousecage 2 pipes 1 incline
9 (IP)	106	"Drop the bucket."	<u>Reduced Constraint:</u> 9 teeter totters 1 boxing glove 5 trampolines 9 ropes 2 pulleys 1 magnifying glass 1 lightbulb 1 dynamite 18 inclines <u>Baseline Constraint:</u> 1 teeter-totter 1 rope 1 pulley 1 magnifying glass 1 lightbulb 1 dynamite

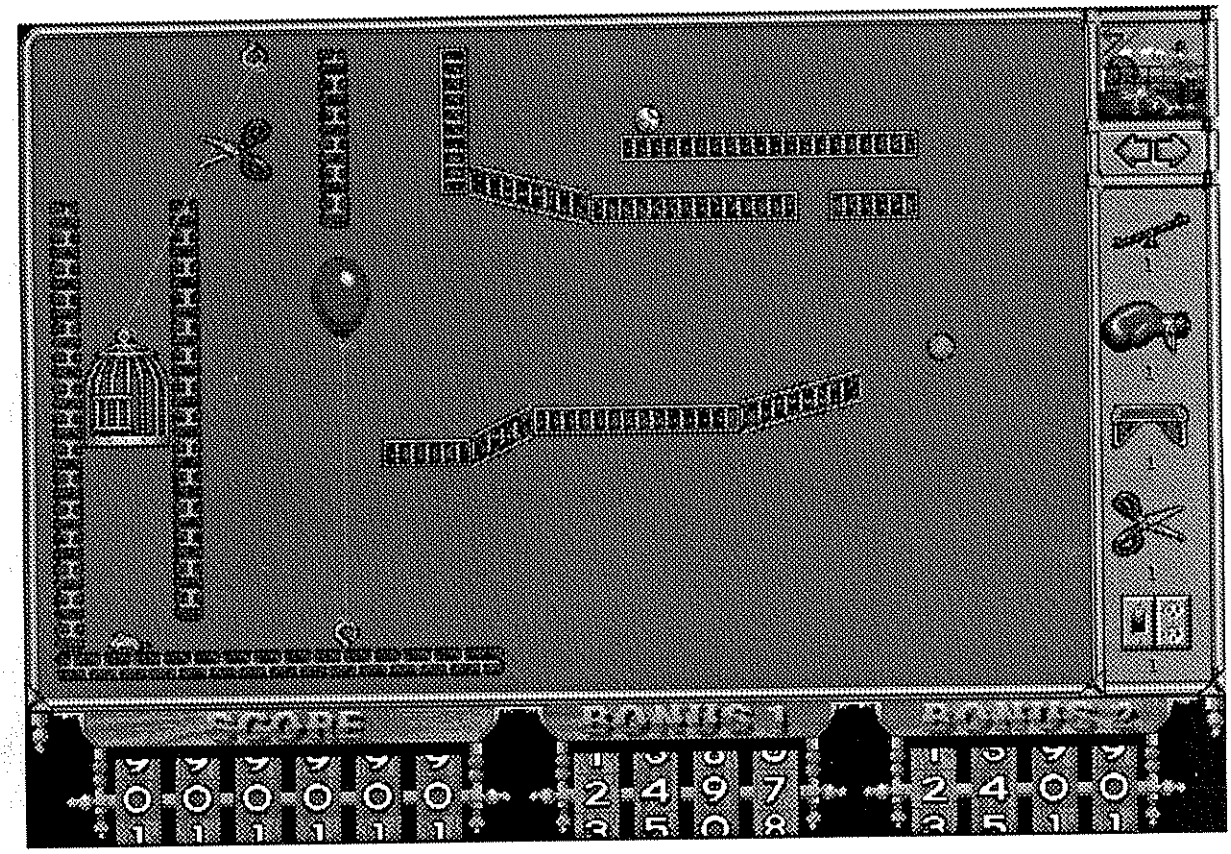
10 (N)	22	"Pop any three of the four ballons. Gears that are turning will pop any balloons that come into contact with them."	4 belts 1 gear 1 rope
11 (IP)	31	"Put all of the basketballs into the wooden container on the right side of the screen."	<u>Reduced Constraint:</u> 1 baseball 2 teeter-totters 1 bellows 1 belt 1 conveyor 1 rope 1 magnifying glass 1 flashlight 1 mousecage 1 incline <u>Baseline Constraint:</u> 1 belt 1 conveyor 1 mousecage 1 incline
12 (N)	121	"Fire off the rocket on the right side of the screen."	2 belts 1 conveyor 1 rope 1 pulley 1 scissors 1 lightbulb 1 candle 1 mousecage

13 (IP)	44	"Make the tennis ball go over the net."	Reduced Constraint: 1 cannon ball 1 teeter-totter 1 trampoline 1 rope 1 pulley 1 gun 1 scissors 1 switch w/outlet 1 fan 1 flashlight 1 dynamite 1 rocket 3 wooden walls 1 incline Baseline Constraint: 1 teeter-totter 1 trampoline 1 rope 1 pulley 1 gun 1 dynamite 1 wooden wall
14 (RS)	94	"Feed Pokey Cat by letting him catch Mort Mouse."	1 belt 2 ropes 3 pulleys 1 scissors 1 magnifying glass 1 lightbulb 1 mousecage 2 inclines
15 (CB)	69	"Turn on all the lights in the house."	5 belts 4 conveyors 4 ropes 2 mousecages 1 incline
16 (CB)	70	"Save Bob the Fish and all his friends. Do not allow any of the fishbowls to be smashed."	1 baseball 1 teeter-totter 1 belt 1 jack-in-the-box 1 scissors 1 mousecage

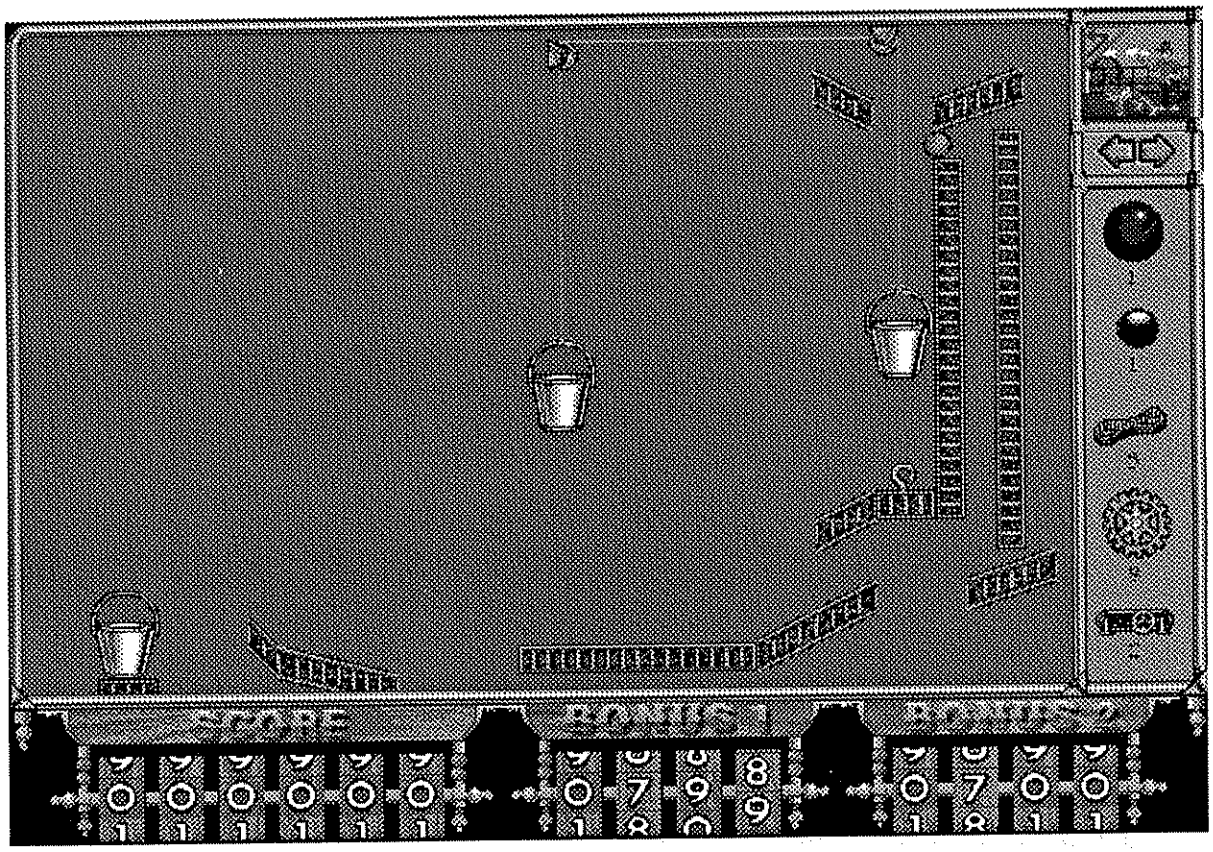
Trial 1



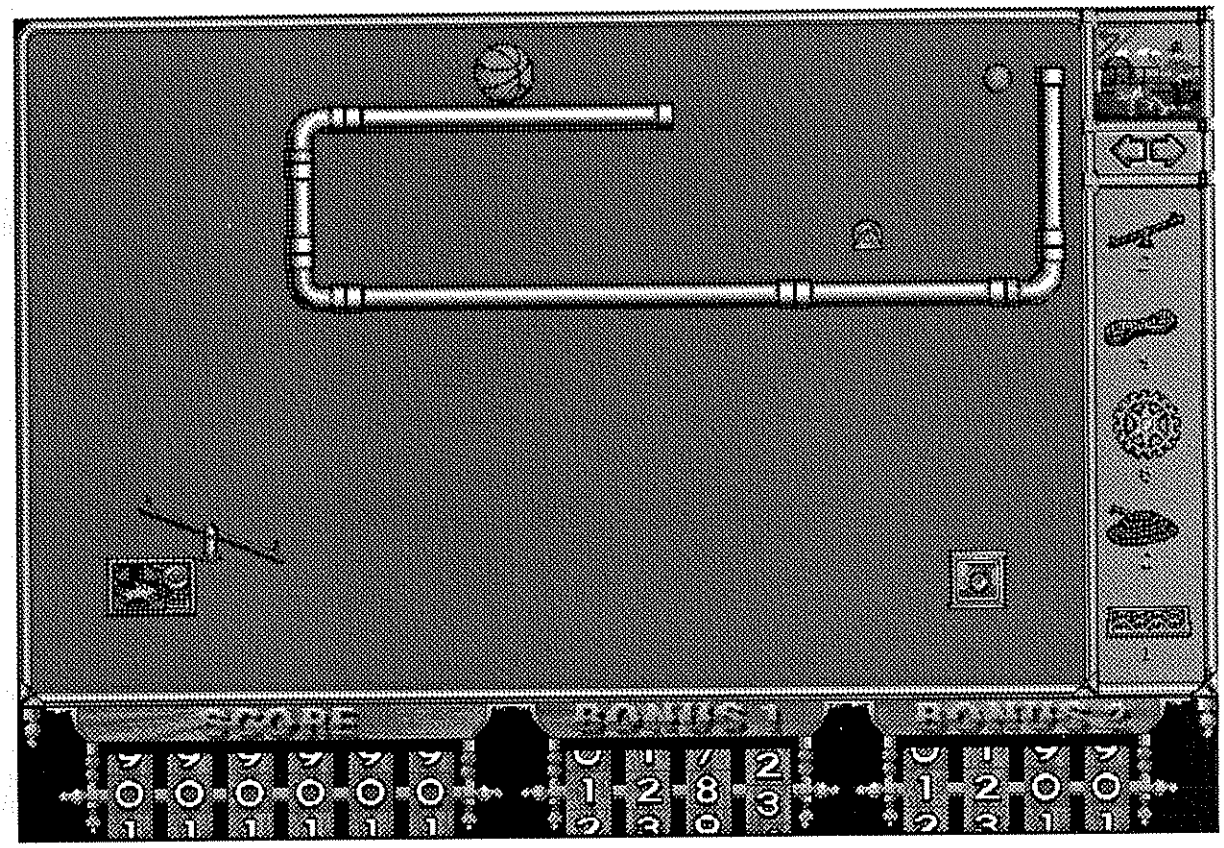
Trial 2



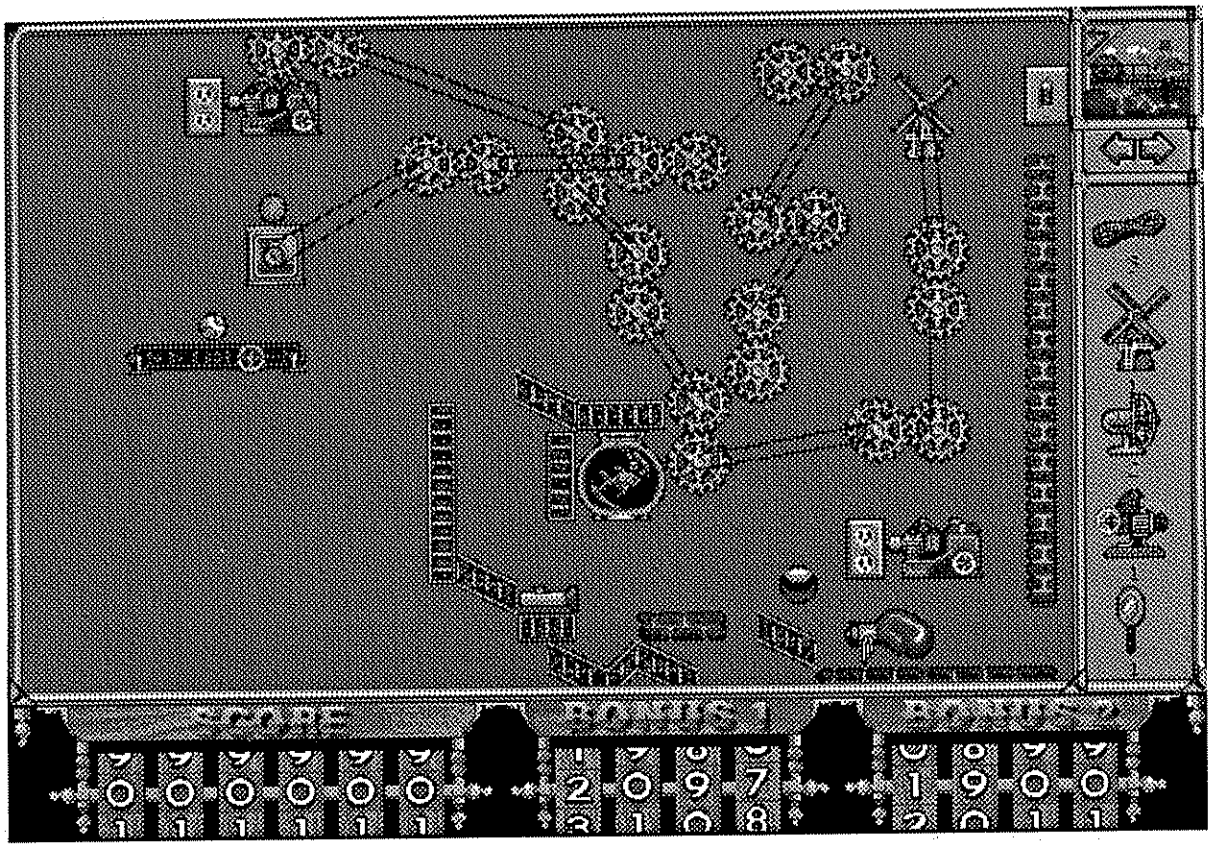
Trial 3



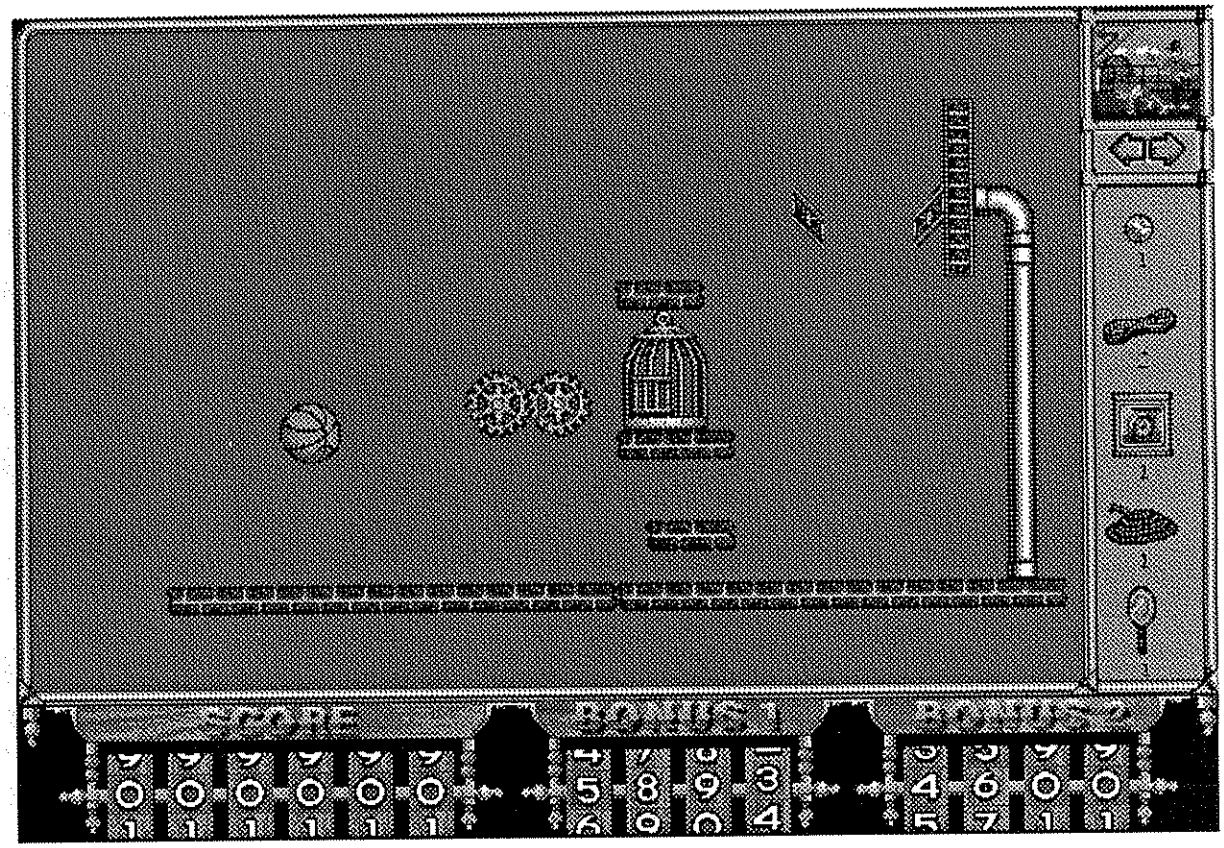
Trial 4



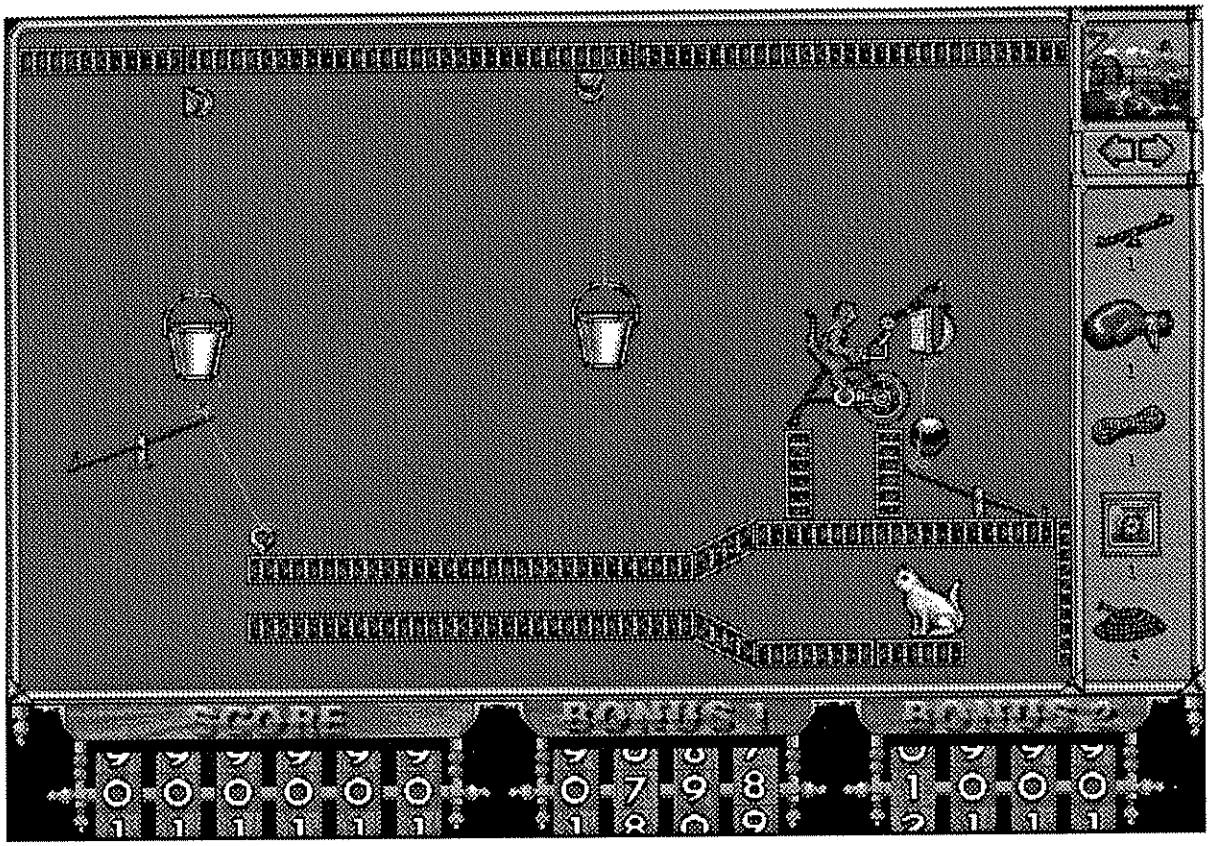
Trial 5



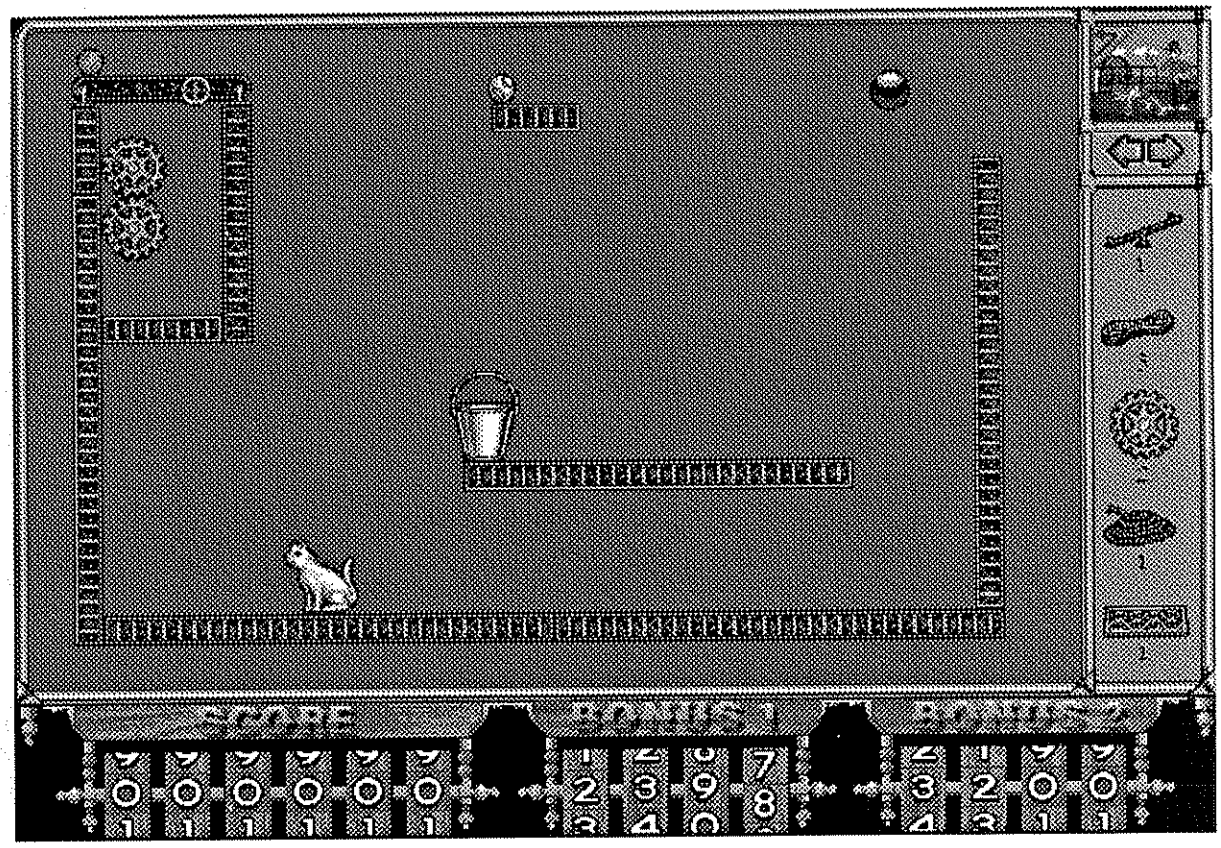
Trial 6



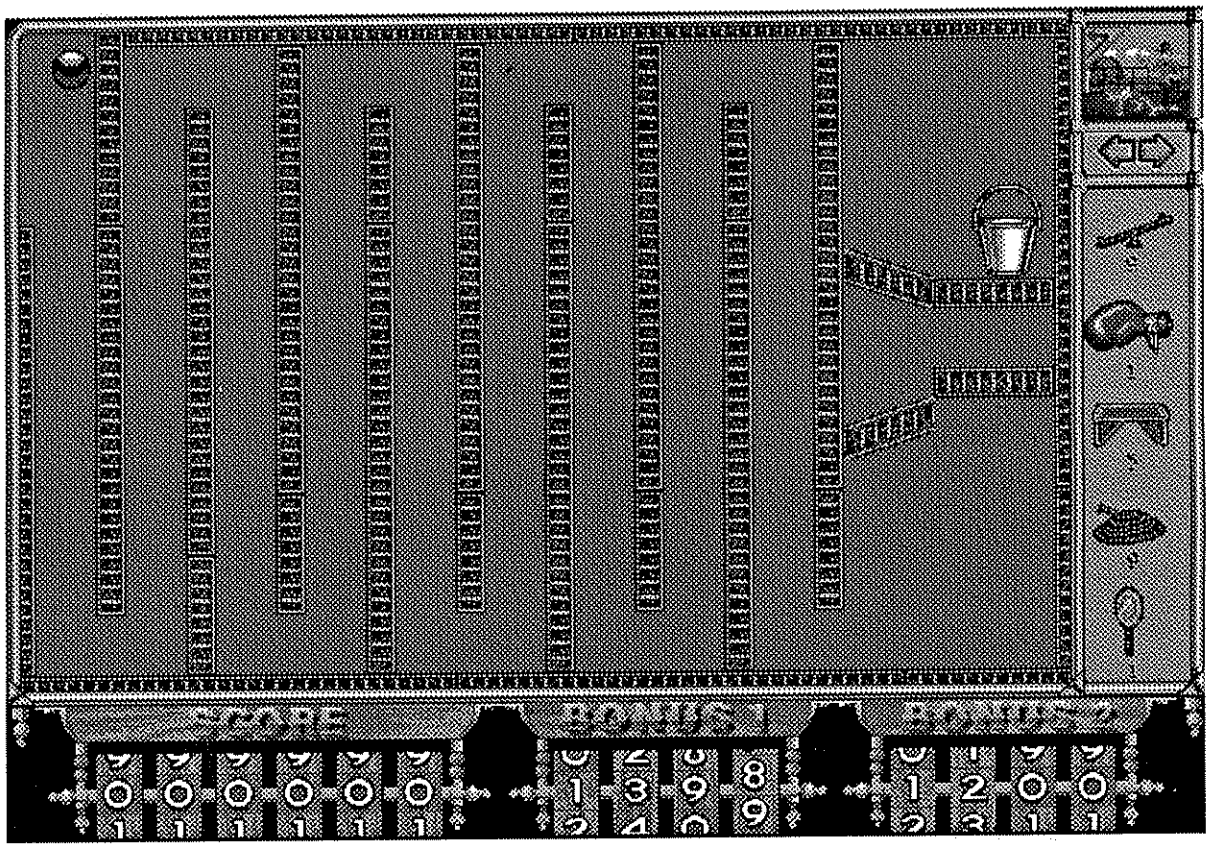
Trial 7



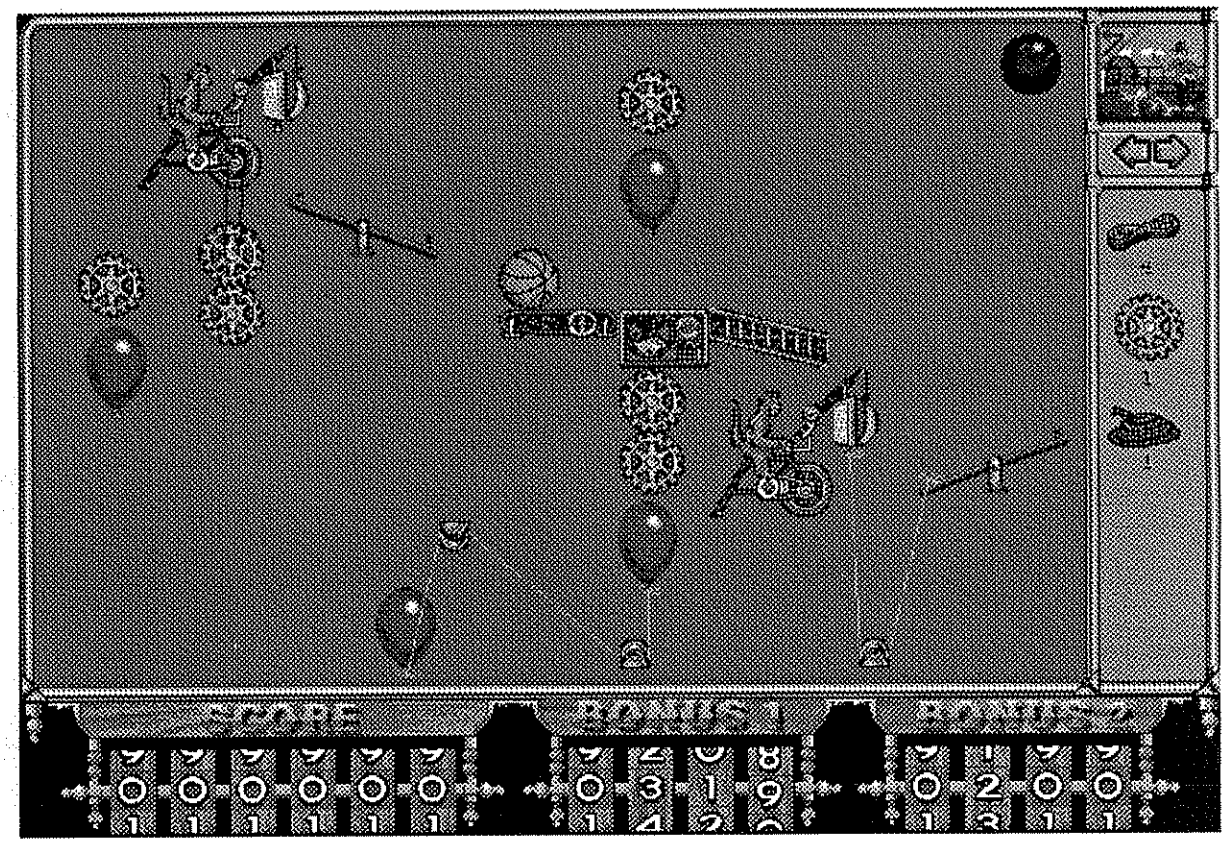
Trial 8



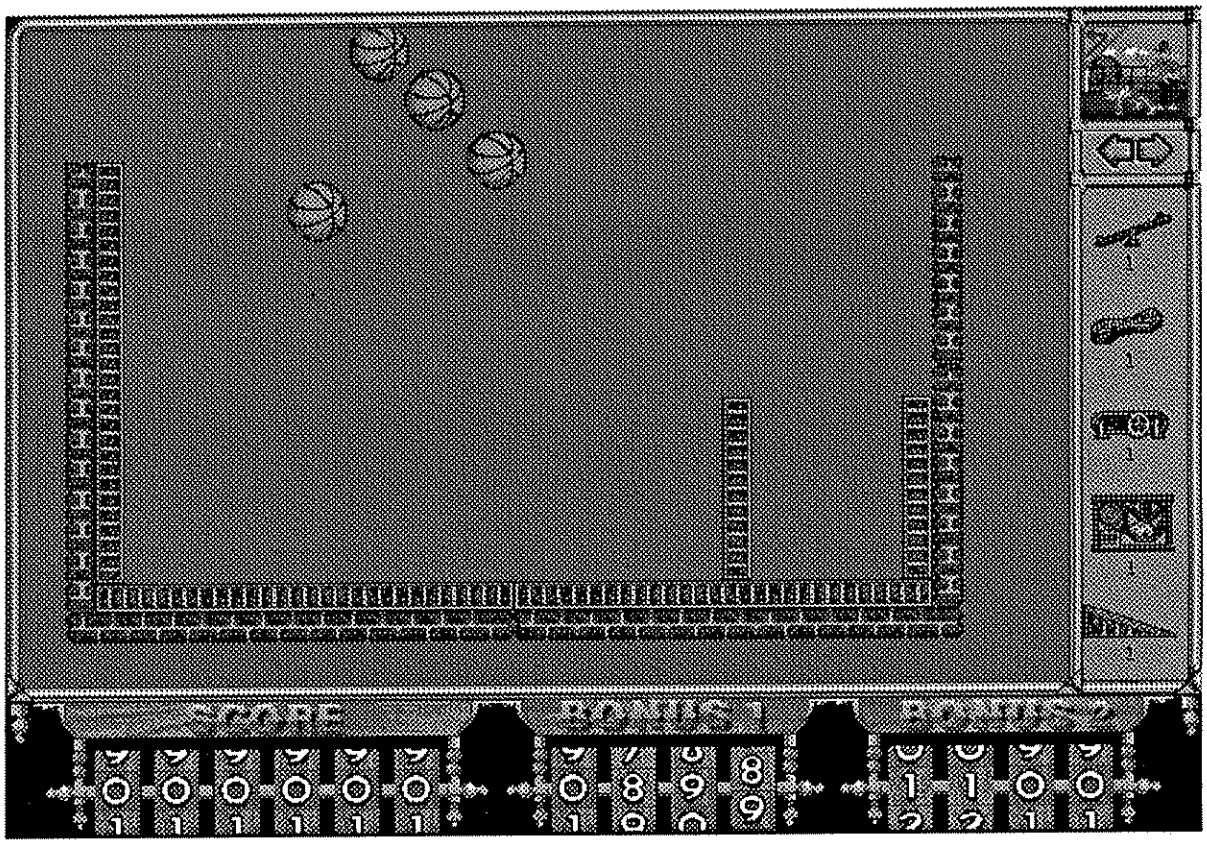
Trial 9



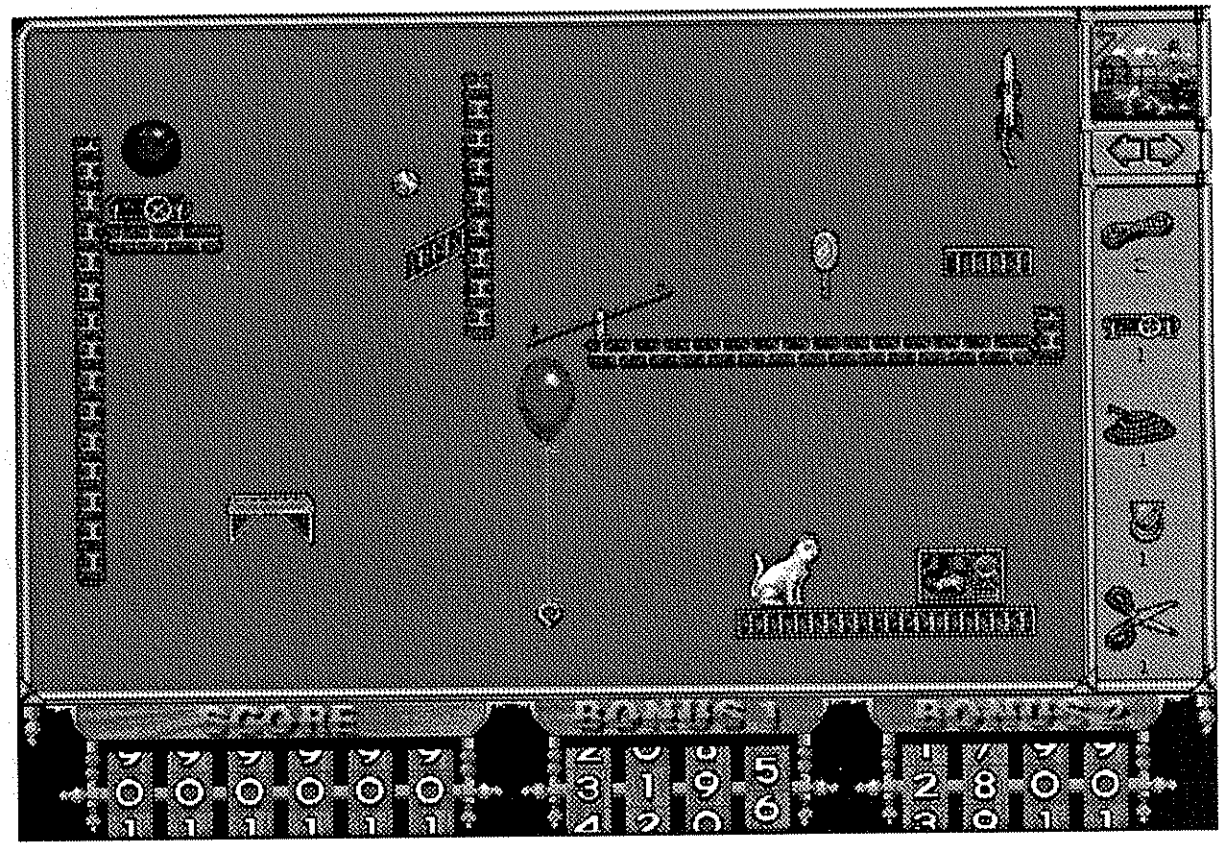
Trial 10



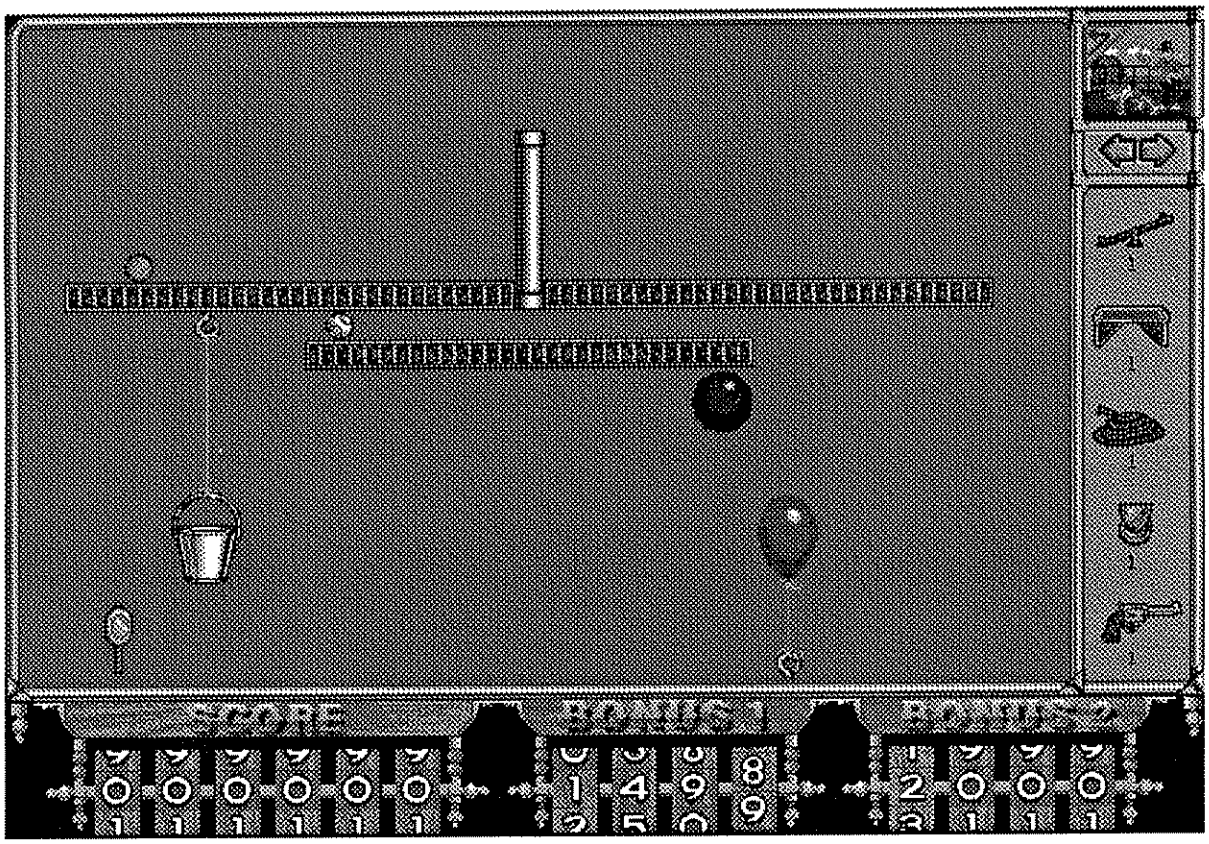
Trial 11



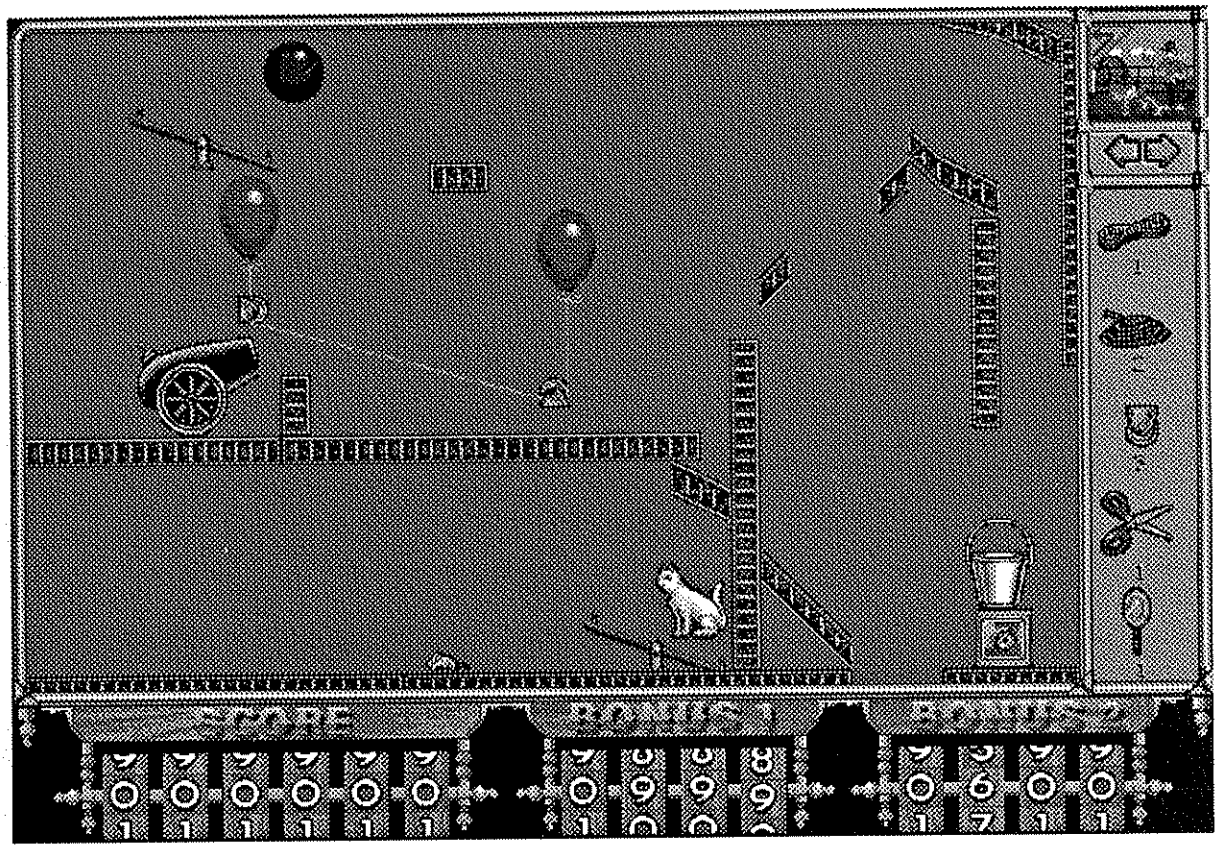
Trial 12



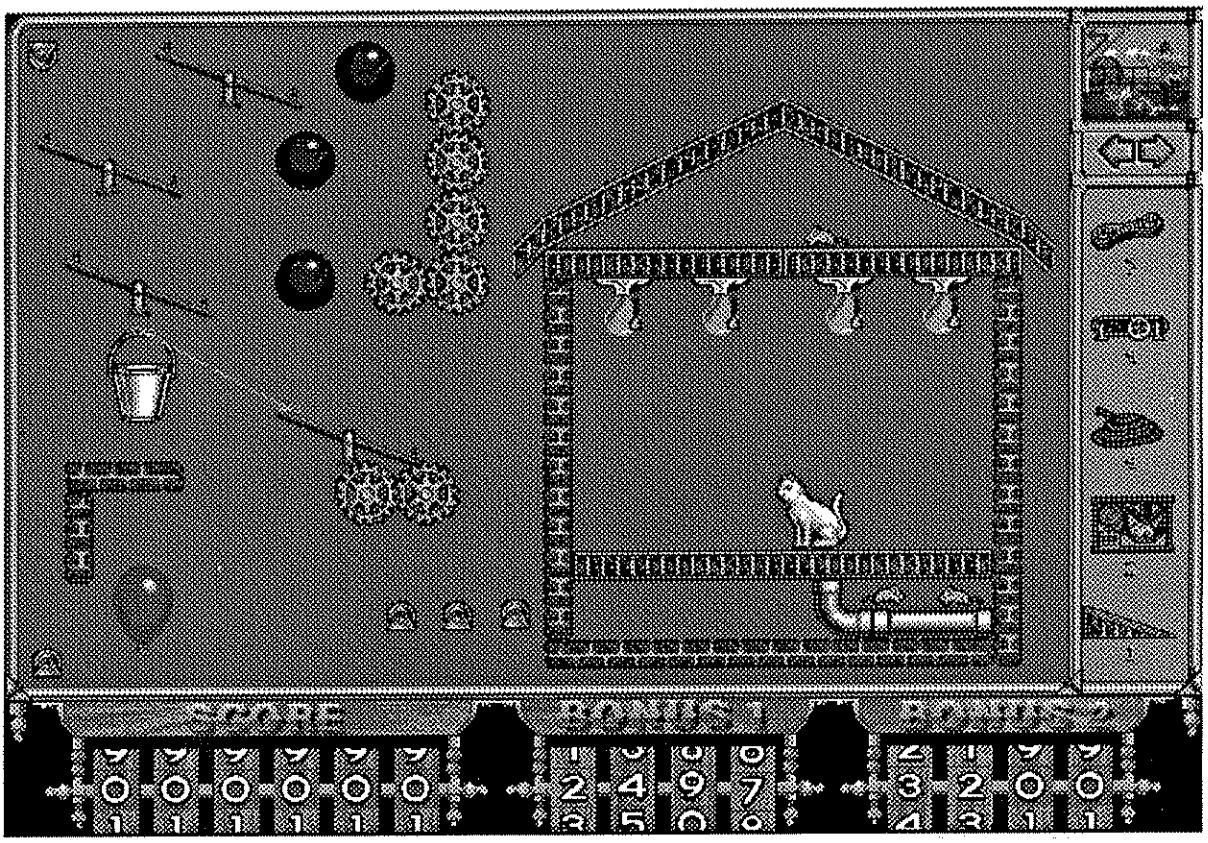
Trial 13



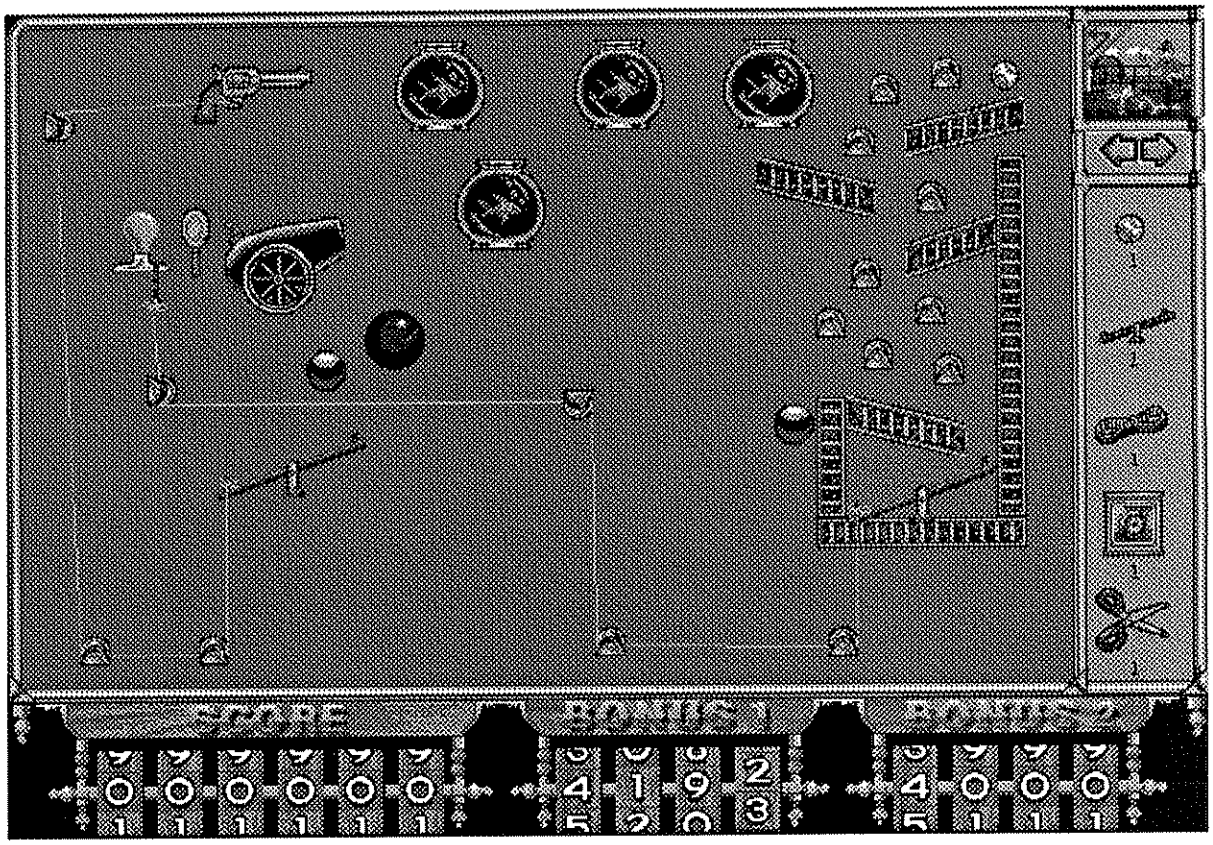
Trial 14



Trial 15



Trial 16



APPENDIX C - Post-Experiment Interview Questions

Baseline Constraint Group

Can you describe your strategy?
- how did you begin each problem?
- did you tend to work backwards from the goal or forwards toward the goal?
- how did you choose which parts to use?

What aspects/types of problems did you find most difficult? Why?

What aspects/types of problems did you find easy? Why?

How did you employ the "start machine" feature to help you?

Did you always try to keep the goal in mind as you solved a problem?

How easily were you able to visualize the behaviour of objects?

Did you find the behaviour of objects intuitive?

Reduced Constraint Group

Can you describe your strategy?
- how did you begin each problem?
- did you tend to work backwards from the goal or forwards toward the goal?
- how did you choose which parts to use?

How did your strategy change on:
- reduced simulation trials?
- covered bin trials?

What aspects/types of problems did you find most difficult? Why?

What aspects/types of problems did you find easy? Why?

How did you employ the "start machine" feature to help you?

Did you always try to keep the goal in mind as you solved a problem?

How easily were you able to visualize the behaviour of objects?

Did you find the behaviour of objects intuitive?

APPENDIX D - Recipe Test Instructions

Subject: Group: Date: Time:

In this procedure, your task is to write out a set of instructions describing how to solve problems in The Incredible Machine (TIM). Don't describe how to solve one particular problem, rather try to describe a general procedure that you might use for approaching problems in TIM. Imagine that the instructions are to be used by someone who has never used TIM before. You may use point form if you wish.