ELSEVIER

Cognition 60 (1996) 205–234

COGNITION

Mental models and temporal reasoning

Walter Schaeken[a,*], P.N. Johnson-Laird[b], Gery d'Ydewalle[a]

[a]*Laboratory of Experimental Psychology, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 , Leuven, Belgium*

[b]*Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA*

Received 24 April 1994, final version accepted 15 December 1995

Abstract

We report five experiments investigating reasoning based on temporal relations, such as: "John takes a shower before he drinks coffee". How individuals make temporal inferences has not been studied hitherto, but we conjectured that they construct mental models of events, and we developed a computer program that reasons in this way.As the program shows, a problem of the form:

a before b
b before c
d while b
e while c
What is the relation between d and e?

where a, b, c, etc. refer to everyday events, calls for just one model, whereas a problem in which the second premise is modified to c before b calls for multiple models because a may occur before c, after c, or at the same time as c.

Experiments 1–3 showed that problems requiring one mental model elicited more correct responses than problems requiring multiple models, which in turn elicited more correct answers than multiple model problems with no valid answers. Experiment 4 contrasted the predictions of the model theory with those based on formal rules of inference; its results corroborated the model theory. Experiment 5 confirmed that a premise leading to multiple models took longer to read than the corresponding premise in one-model problems, and that latency to respond correctly was greater for multiple-model problems than for one-model problems. We conclude that the experiments corroborate the mental model theory.

* Corresponding author. Fax: 016 28 60 99.

PII S0010-0277(96)00708-1

1. Introduction

Imagine that the following facts have been established beyond a reasonable doubt:

> After the suspect ran away, the clerk rang the alarm.
> The manager in the bank was stabbed while the alarm was ringing.
> What is the temporal relation between the suspect running away and the stabbing of the manager?

Most people are likely to infer that the suspect ran away before the stabbing of the manager. This conclusion is valid, that is, it must be true given that the premises are true. Such inferences, which hinge on the temporal relations between events, are ubiquitous and often important in daily life.

Psychologists have studied the perception of time (e.g., Fraisse, 1963), the development of the concept of time (e.g., Piaget, 1969), and the psycholinguistics of temporal expressions (e.g., Miller and Johnson-Laird, 1976, Sec. 6.2). They have studied the role of time in causal reasoning (e.g., Girotto et al., 1991), and the comprehension of temporal descriptions (e.g., Oakhill and Garnham, 1985; Mandler, 1986), but they do not appear to have investigated temporal reasoning itself. How individuals make such inferences is unknown, but the present paper aims to explain this ability.

A common view in cognitive science is that reasoning depends on mental rules of inference akin to those of a logical calculus (see, for example, Braine et al., 1984; Macnamara, 1986; Osherson, 1975; Pollock, 1989; Rips, 1983; Rips, 1994). In contrast, our hypothesis is that individuals use their knowledge of the language and their general knowledge to construct mental models of temporal sequences of events. This idea relates to studies of relational reasoning (e.g., Huttenlocher, 1968), to studies of spatial reasoning (e.g., Johnson-Laird, 1983; Byrne and Johnson-Laird, 1989), and to various algorithms in artificial intelligence that construct temporal models on the basis of verbal descriptions.

Thus, for example, Isard (1974), building on earlier work in Isard and Longuet-Higgins (1971) and ultimately on Reichenbach's (Reichenbach, 1947) analysis of tense, developed a program that answered questions about games of tic-tac-toe which it played with a human user, and the program was sensitive to tense, mood, aspect, auxiliary verbs, if-clauses, and when-clauses. Steedman (1982) also developed a program that simulates a multi-user operating system and that shows how to answer temporal questions about the users of the system. He argued that the representation of temporal events should be organized along a time line, which encodes the relative time of events and the start and end of intervals of time corresponding to such assertions as "while you talked to Mary". We now outline the mental model theory of temporal reasoning, which was inspired in part by these and other exercises in artificial intelligence (see, for example, Allen, 1983; Dinsmore, 1991).

Deductive reasoning, according to the theory of mental models, depends on constructing a set of models based on the premises, formulating a conclusion if none is provided, and ensuring that no model of the premises falsifies it (Johnson-Laird, 1983). If a conclusion is true in all the models of the premises, then it is necessary (valid); if it is true in most of the models of the premises, then it is probable; if it is true is at least some model of the premises, then it is possible; if it is true in only a few models of the premises, then it is improbable; and if it is true in none of the models of the premises, then it is impossible (inconsistent with the premises). The central representational assumption of the theory is that a model corresponds to a set of possible situations, which it represents by encoding those aspects of their structure that they have in common. Models accordingly represent the structure of situations rather than the logical form of premises (see Johnson-Laird and Byrne, 1991, p. 38).

The theory makes two main predictions. First, the more alternative models that have to be constructed in order to draw a correct conclusion, the longer the task should take, and the greater the chance of error should be. Second, erroneous conclusions should tend to be consistent with the premises rather than inconsistent with them. Such conclusions arise because reasoners are likely to base them on at least some model of the premises but overlook other possible models. This prediction can be tested in the absence of a detailed account of the numbers or sorts of models yielded by premises: one merely has to assess whether an erroneous conclusion could be true given the truth of the premises.

Both predictions have been corroborated by previous experiments in the major domains of deduction, including reasoning with propositional connectives, quantifiers, and relational expressions (see, for example, Johnson-Laird and Byrne, 1991; Johnson-Laird et al., 1992). Neither prediction, however, is easy to derive from existing accounts based on formal rules (e.g., Braine et al., 1984; Hagert, 1984; Ohlsson, 1984; Osherson, 1975; Rips, 1994).

Temporal relations are unlikely to be visualized in a single static image. The events themselves may not be visualizable, and indeed manipulations of imageability have no detectable effects on reasoning (see, for example, Newstead et al., 1982; Richardson, 1987; and Johnson-Laird et al., 1989). Mental models, however, can represent situations that are not visualizable, and if subjects reason about temporal relations by constructing models, then two obvious sort of models are open to them. One sort of model of a temporal sequence could itself unfold in time kinematically, though not necessarily at the same speed as the original events themselves. This sort of representation uses time itself to represent time (see Johnson-Laird, 1983, p. 10). A second sort of model represents temporal relations statically as a sequence of events akin to a spatial model except that the main axis corresponds to time. The various sorts of temporal relation, at least as expressed in English (see, for example, Allen, 1983), can all be represented spatially, and thus according to this account temporal reasoning depends on mapping spatial expressions into static models in which one dimension represents time. Temporal relations might be slightly harder to represent than spatial relations, because they

have to be transformed from a temporal to a quasi-spatial medium. We shall return to the difference between static and kinematic models later in the paper, but now we turn to an algorithm for temporal reasoning.

2. A model-based algorithm for temporal reasoning

The model theory specifies an interpretative process that leads from premises to models, and a descriptive process that leads from models to conclusions. The first stage of the interpretative process is a "compositional semantics", which specifies how the mind constructs a semantic representation of the meaning of sentences. The theory assumes that each lexical entry contains information about the word's contribution to the truth conditions of assertions, that each grammatical rule has a related semantical rule, and that the parser uses this information to combine the meanings of constituents according to the grammatical relations amongst them (see Johnson-Laird, 1983, for the theory's account of the mental lexicon and grammar, and the design of the mental parser). The particular proposition that a sentence expresses also depends on general and contextual knowledge. We shall simplify by treating context as the information that is already represented in the model(s) of the discourse so far. The second stage of the interpretative process uses the semantic representation of a sentence to update any model of the situation or to build one ab initio. The semantic representation and the existing models, if any, are used to determine which procedure should be used to update the models (see the account of the program below). Finally, the descriptive process leads to the construction of a conclusion if none exists. Conclusions are formulated by scanning the models for a parsimonious and novel relation. The theory assumes that reasoners attempt to construct all possible models as they interpret each of the premises in the order in which they are stated. But, if the number of possibilities grows too large for the capacity of working memory, they can adopt a procedure that allows them to ignore any irrelevant premises. In certain domains, particularly those depending on quantifiers, the theory proposes that not all of the possible models are obvious, and so reasoners base a conclusion on an initial model and then test its validity by searching for alternative models. This account provides an alternative explanation for the so-called "atmosphere" effect in syllogistic reasoning (see Johnson-Laird and Byrne, 1991). In other domains, notably reasoning with sentential connectives, the theory yields initial models that in certain cases yield systematically erroneous conclusions – a prediction that has also been corroborated (see Johnson-Laird and Savary, 1995).

The algorithm for temporal reasoning uses static models. The representation of the assertion:

The clerk sounded the alarm after the suspect ran away

thus calls for a model of the form:

r a

in which the time axis runs from left to right, "r" denotes a model of the suspect running away, and "a" denotes a model of the clerk sounding the alarm. Events can be described as momentary or as having durations, definite or indefinite. Hence, the further assertion:

The manager was stabbed while the alarm was ringing

means that the stabbing occurred at some time between the onset and offset of the alarm:

r a——

s

where "s" denotes a model of the stabbing. This model corresponds to infinitely many different situations that have in common only the truth of the two premises. For example, the model contains no explicit representation of the duration for which the alarm sounded, or of the precise point at which the stabbing occurred. Yet, the conclusion:

The stabbing occurred after the suspect ran away

is true in this model, and there is no model of the premises that falsifies this conclusion.

We have implemented a computer program in LISP that carries out temporal inferences in the same way. The inferences include all of those that we used in our experiments. The program has a context-free grammar for a fragment of English that contains premises of the form, "a happens before b", "b happens while c", and so on. Its compositional semantics constructs a semantic representation of any sentence in the fragment based on a representation of lexical meanings and on semantic rules associated with each rule in the grammar. This semantic representation is used to update the set of models. Given the semantic representation of the assertion:

a happens before b

the program checks whether a or b, or both, are already represented in any model of the discourse. If neither occurs in an existing model, the program starts a new model; if one but not the other occurs in any existing model, then the model is updated to include the new event referred to in the premise; if one occurs in an existing model and the other occurs in a different existing model, then a new model, or set of models, is formed by combining these models according to the premise; and if both occur in the same models, then the truth of the premise is

checked in these models. Because the program constructs all possible models of co-referential premises, the process of verifying an assertion yields one of the following responses: the assertion is a valid deduction (it is true in all the models of the premises), it was previously possibly false (it is false in some of the models, which are duly eliminated), or it is inconsistent with the previous premises (i.e., it is false in all the models). Given a question about the relation between two events, the program formulates a conclusion if a common relation holds between them over all the models of the premises; otherwise, it responds that there is no definite relation between the two events, either because different relations occur in different models or because the events do not occur in any one model.

For simplicity, the program does not represent the relative durations of events: in effect, it assumes that they are all of the roughly the same duration. Thus, for example, given the following problem:

a happens before b
b happens before c
d happens while a
e happens while c
What is the relation between d and e?

the program constructs a model corresponding to the following array with time running from left to right:

a	b	c
d		e

from which it formulates the answer that d happens before e. Like the premises in this problem, this model corresponds to infinitely many possible situations depending on the actual onsets and offsets of each event.

Some temporal descriptions are radically indeterminate. For example, the following two premises:

a happens before b
c happens before b

do not fix the temporal order of a and c. In such cases, the program constructs models corresponding to the three possibilities: a happens before c, c happens before a, a and c happen contemporaneously. The construction of all possible models is, in principle, intractable: it yields an exponential growth in the number of models as indeterminacies in the premises mount up. Yet, the procedure is feasible as long as there is only a small number of indeterminacies, and there is evidence that reasoners do try to keep track of alternative models, at least in those domains of reasoning that do not depend on quantifiers, that is, in reasoning that hinges only on simple relations (Byrne and Johnson-Laird, 1989), or on sentential connectives (Johnson-Laird et al., 1992). Human performance, as we will see,

rapidly degrades with an increasing number of models – a phenomenon that is predictable assuming that the human inferential system uses an intractable algorithm and a working memory of limited capacity.

Certain problems yield many possible models and yet seem to be solved rapidly. For example, with the following problem:

a happens before e
b happens before e
c happens before e
d happens before e
What is the relation between a and d?

one readily grasps that there is no definite relation between a and d. It is easy to see, however, that the only relevant premises are the first and the last because only they refer to the events in the question. They yield the models:

a d e

and:

d a e

which establish that there is no definite relation between a and d. The rest of the premises are irrelevant to the question, and so there is no need to use them to construct models. Hence, when reasoners have immediate access to all the premises and the question, they can construct models from just those premises that are relevant to its answer. Where the premises are not co-referential, as in:

a happens while b
c happens while d
e happens while f
g happens while h
What is the relation between a and g?

each premise has a separate model, and the models are not combined because they are not co-referential (for evidence supporting this principle, see Ehrlich and Johnson-Laird, 1982). Hence, there is clearly no definite relation between a and g.

We have implemented these ideas in the program in the following way. The program is restricted in the number of models that it can construct in trying to solve a problem (by analogy with a limited capacity working memory). When the models it has constructed exceed this number, it then searches for a chain of premises interrelating the two events in the question, and constructs models only from them. If there is no question, then the program cannot use this strategy. As an example of the strategy, consider the following premises:

k happens before a
a happens before b
h happens before b
b happens before i
b happens before c
c happens before d
e happens before d
f happens before d
d happens before g
What's the relation between a and d?

When the program works through the premises in their stated order, it constructs 2347 models for the first eight premises – a number that vastly exceeds the capacity of human working memory. If the program's capacity is set more plausibly, say, to four models, it will give up working forwards and then try a depth-first search based on the question. Fig. 1 shows the referential structure of the premises (not a model of them), and the program discovers the co-referential chain of premises:

a happens before b
b happens before c
c happens before d

leading from a to d. It constructs the single model that these premises support, which yields the conclusion:

a happens before d

The advantages of this procedure are two-fold. First, it ignores all irrelevant premises that are not part of the chain connecting one event in the question to the other. Second, it deals with the premises in a co-referential order in which each premise after the first refers to an event already represented in the set of models. In an unpublished study, we have shown that people do appear to learn to ignore irrelevant premises when the question to be answered is posed before the presentation of the premises.

In everyday life, speakers are likely to present information in an amount and an

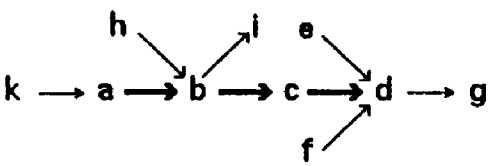

Fig. 1. The referential structure of a set of premises, where each arrow denotes a premise relating the two events it connects. To answer the question, "what is the relation between a and d?", the program discovers the direct chain of premises shown here by the heavier arrows.

order that does not overburden human working memory; they are likely to be sensitive to the limitations of their audience (see Grice, 1975). Hence, in the first study of temporal reasoning we used similarly straightforward materials. We can therefore base our predictions on the number of models that the program constructs working through the premises in their stated order. Indeed, as we shall see, there is evidence that our subjects worked in this direction too. Our experiments, as we explain in the next section, investigated three main sorts of problem: one-model problems, multiple-model problems with valid answers, and multiple-model problems with no valid answers.

EXPERIMENTS 1–3

The aim of our initial experiments was to test the two main predictions of the model theory in the domain of temporal reasoning: an inference that depends on one model should yield fewer errors than one that depends on multiple models, and erroneous answers should be consistent with the premises. These experiments also allow us to examine the predictions in cases that obviate Rips's (Rips, 1994) criticism that an instruction to imagine objects on a table-top biases subjects to respond in ways that favor models. The present experiments used no such instruction. Experiment 1 was an exploratory study using school children as subjects. We had to reject the data from many of them, and so Experiment 2 was a replication with university students as subjects, and Experiment 3 used an improved design also with university students. Each of the experiments examined three sorts of temporal deduction:

1. One-model problems; that is, those in which the premises yield only one model and so are bound to yield a valid answer.
2. Multiple-model problems with a valid answer; that is, those in which the premises yield multiple models that all support the same answer, which is accordingly valid.
3. Multiple-model problems with no valid answer; that is, those in which the premises yield multiple models that do not support an answer in common.

The problems concerned the temporal sequences of events, and were based on the connectives "before", "after", and "while" ("voordat", "nadat", and "terwijl" in Flemish, which was the language for all of our experiments). We will give just one example of each sort of problem because the temporal connectives were systematically manipulated as were the specific events denoted here by letters of the alphabet. A one-model problem is illustrated by the premises:

1. a happens before b
   b happens before c
   d happens while b
   e happens while c
   What is the relation between d and e?

This problem yields the model:

a b c

d e

and this model supports the answer:

d happens before e

The premises do not support any model that refutes this answer, and so it is valid; that is, it must be true given that the premises are true.

A multiple-model problem with a valid answer is illustrated by the following premises, which differ from the one-model problems in the order of terms in the second premise; that is, it has the same events and the same temporal relation ("before"):

2. a happens before b
c happens before b
d happens while c
e happens while b
What is the relation between d and e?

This problem yields at least two alternative models:

a c b c a b

d e d e

In principle, it also yields a model in which a and c occur at the same time (as our computer program demonstrates), but for simplicity we will ignore this possibility henceforth. These models of the premises support the answer:

d happens before eand so it is valid.

A multiple-model problem with no valid answer is illustrated by the following premises, which differ from the previous problem only in one term of the final premise; that is, it has the same temporal relation ("while"):

3. a happens before b
c happens before b
d happens while c
e happens while a
What is the relation between d and e?

This problem yields at least two alternative models:

a c b c a b

e d d e

but they do not support any answer in common about the relation between d and e. Hence, there is no valid answer about this relation. Readers will note that the first premise is irrelevant in the first two sorts of problem, but not in this third sort of problem. We will return to this point later.

The model theory predicts that the greater the number of models that have to be constructed the more likely errors are to occur. One-model problems are bound to have a valid answer, whereas multiple-model problems may, or may not, have one. Reasoners who construct only one of the models of a multiple-model problem with a valid answer can nevertheless draw the correct answer, because all the models support the same answer. However, subjects who construct only one of the models of a multiple-model problem with no valid answer will draw a answer where none is warranted. Hence, the construction of the full set of models is more critical for the multiple-model problems with no valid answers than for those with valid answers. The first prediction of the model theory is accordingly the following trend in increasing errors: one-model problems, multiple-model problems with valid answers, and multiple-model problems with no valid answers. The nature of the correct responses differs in the last case; that is, a conclusion should be drawn for the problems with valid answers but no conclusion should be drawn for the problems with no valid answers. A second prediction, where the correct responses are the same, is that one-model problems should be easier than multiple-model problems with valid answers.

Experiment 1 also examined two sorts of deduction based only on two premises: one-model problems with valid answers, and multiple-model problems with no valid answers. With just two premises, there are no multiple-model problems that support valid deductions apart from those that merely re-state the information provided by one of the premises. The one-model problems are illustrated by the following premises:

4. a happens before b
   b happens before c
   What is the relation between a and c?

   The problem yields the model:

   a b c

   which supports the valid answer:

   a happens before c

   The multiple-model problems with no valid answers are illustrated by the following premises:

5. a happens before b
   a happens before c

What is the relation between b and c?

The problem yields at least the following two models:

a b c a c b

which do not support any valid answer relating b and c. If number of models is the critical factor, then there should be no marked difference between the two-premise and four-premise problems provided that the size of the individual models themselves is not too large for working memory, i.e., there is little difference between a model of three events and a model of five events (see also Byrne and Johnson-Laird, 1989).

3. Method

3.1. Design

In all three experiments, the subjects acted as their own controls and carried out: (1) one-model problems based on four premises with a valid answer; (2) multiple-model problems based on four premises with a valid answer; and (3) multiple-model problems based on four premises with no valid answer. Table 1 summarizes these problems. In Experiment 1 only, the subjects also carried out: (4) one-model problems based on two premises with a valid answer; and (5) multiple-model problems based on two premises with no valid answer.

In Experiments 1 and 2, each subject carried out four inferences for each of the different sorts of problem presented with a different content, and in Experiment 3 each subject carried out eight inferences for each of the different sorts of problem, that is, totals of 20 problems in Experiment 1, 12 problems in Experiment 2, and 24 problems in Experiment 3. The order of presentation was randomized for each subject.

3.2. Materials

Four versions of each problem were constructed in the following way: the temporal relations ("before", "after") in the first and second premises were systematically manipulated so that in the first version both premises contained "before", in the second version the first premise contained "before" and the second premise contained "after", in the third version the first premise contained "after" and the second premise contained "before", and in the fourth version both premises contained "after". The remaining premises in the problems were held constant. Once the relational terms are fixed in this way, the arrangement of the terms in the problems is also fixed given the particular sort of problem. In the first two experiments, where there was a valid answer, the event referred to by the first

Table 1
A summary of the the three sorts of problem used in Experiments 1–3 and the additional sort of one-model problem used in Experiments 4 and 5. The problems are shown with illustrative examples and diagrams of their models (omitting the models where a and c happen contemporaneously in the multiple-model cases)

One-model problem with irrelevant first premise						
a happens before b						
b happens before c	a	b	c			
d happens while b		d	e			
e happens while c						
What is the relation between d and e?						
One-model problem with transitive relation (used only in Expts. 4, 5)						
a happens before b						
b happens before c	a	b	c			
d happens while a	d		e			
e happens while c						
What is the relation between d and e?						
Multiple-model problem with valid answer						
a happens before b						
c happens before b	a	c	b	c	a	b
d happens while c		d	e	d		e
e happens while b						
What is the relation between d and e?						
Multiple-model problem with no valid answer						
a happens before b						
c happens before b	a	c	b	c	a	b
d happens while c	e	d		d	e	
e happens while a						
What is the relation between d and e?						

item in the question occurred before the event referred to by the second item in the question. This constraint should tend, if anything, to reduce the difference between one-model and multiple-model problems, but we eliminated it in Experiment 3 by using two different variants of each of the four versions of a problem; that is, we manipulated the order of the third and fourth premises and the order of the two terms in the final question. Thus, corresponding to the one-model problem above, we also used the following version:

a happens before b
b happens before c
e happens while c
d happens while b
What is the relation between e and d?

In this case, the event referred to by the first item in the question occurred after the event referred to by the second item in the question. The two variants of each

version yielded eight distinct ways of stating a problem.The three experiments and the subsequent ones were carried out in Leuven, Belgium and the materials were in Flemish (Dutch). The problems were based on two sorts of lexical materials. The first sort concerned the temporal order of cartoons on television networks; for example:

> The cartoon "the tiny tot" is on channel 1 before the cartoon "the mighty mouse" is on channel 1.

The premises containing "while" referred to cartoons on channel 2. The cartoon names consisted of definite descriptions (article, adjective, noun) in which the adjective and noun began with the same consonant in Flemish in order to help the subjects to remember them. Each of the problems had a different set of names, which were assigned at random from a pool of 84 names.

The second sort of lexical materials concerned the temporal order of everyday activities; for example:

> John takes a shower before he drinks coffee.

Each problem was based on a separate set of activities selected at random from a pool of such events, and the names were selected at random from a pool of 16 female and 16 male first names.

In Experiments 1 and 2, the materials were assigned to three groups of subjects in the following way. One group received the cartoon materials with the main clause prior to the subordinate clause, for example, a before b. A second group received the cartoon materials with the subordinate clause prior to the main clause, for example, before b, a. A third group received the everyday activities with the main clause prior to the subordinate clause. In Experiment 3, there were only two sorts of materials: one group of subjects received the cartoon materials and another group received the everyday activities, and for both groups the main clause was prior to the subordinate clause.

3.3. Procedure

In each experiment, we tested the subjects in a single group. The instructions of the task were written down on the first page of a booklet given to each subject. They explained that the subjects' task was to answer a series of questions based on the information about the order of events given in the preceding assertions, and that the answers should be those that *must* be true given the truth of the previous assertions. If the subjects thought that there was no definite answer, they had to write that down as their response. The experimenter then gave the subjects one practise problem, which was a two-premise problem with a valid answer. Each problem, together with the question, was printed on a separate page in the booklet, and the subjects had to write the answer under the question. They were asked not

to return to a question once they had answered it. The experiments lasted from 25 to 40 minutes depending on the number of problems.

3.4. Subjects

Seventy-two subjects aged from 17 to 19 from the last year of a secondary school participated voluntarily in Experiment 1. Fifty-two subjects participated in Experiment 2. They were all first-year psychology students, who were fulfilling a course requirement. Thirty-six subjects from the same population participated in Experiment 3.

4. Results and discussion

We had to discard the data from 26 of the school children in Experiment 1, because they did not answer all the questions, but we had to discard the data of only three and four of the adult subjects from Experiments 2 and 3, respectively. The percentages of correct responses for the remaining subjects in the three experiments are shown in Table 2. There were no significant differences between the different sorts of material in any of the three experiments, and so we have pooled the results.

For the three sorts of problem common to all the experiments, that is, those based on four premises, the predicted trend in errors was corroborated: one-model problems yielded fewer errors than multiple-model problems with valid answers, which yielded fewer errors than multiple model problems with no valid answers. A non-parametric trend test devised by Page (1963) for repeated measures was significant both by subjects (for Experiment 1, Page's $L = 587$, $n = 46$, $p < .0001$; for Experiment 2, $L = 628$, $n = 49$, $p < .00003$; and for Experiment 3, Page's $L = 402$, $n = 32$, $p < .02$) and by materials (for Experiments 1 and 2, Page's $L = 56$, $n = 4$, $p < .001$; and for Experiment 3, Page's $L = 109$, $n = 8$, $p < .001$).

Table 2
The percentages of correct responses in all five experiments. The one-model problems either have an irrelevant premise (irrel. premise) or else contain a transitive relation (trans. rel.); the multiple-model premises either have a valid answer or else have no valid answer (n.v.)

Types of problem						
	Two premises		Four premises			
	One-model	Multiple-model (n.v.)	One-model (irrel. prem.)	One-model (trans. rel.)	Multiple-model	Multiple-model (n.v.)
Expt. 1	93	64	93	–	86	60
Expt. 2	–	–	92	–	81	65
Expt. 3	–	–	91	–	87	67
Expt. 4	–	–	94	–	84	19
Expt. 5	–	–	89	93	81	44

Likewise, the one-model problems yielded fewer errors than the multiple-model problems with valid answers (for Experiment 1, Wilcoxon's $T = 82$, $n = 13$, $p < .005$; for Experiment 2, Wilcoxon's $T = 195.5$, $n = 21$, $p < .003$). This difference, however, was not significant in Experiment 3, though the trend was in the predicted direction (Wilcoxon's $T = 93.5$, $n = 16$, $p > .09$). In Experiment 1, there was no reliable difference between the two-premise and four-premise problems (Wilcoxon's $T = 236$, $n = 34$, $p > .15$), but the one-model problems yielded fewer errors than the multiple-model problem with no valid answer (Wilcoxon's $T = 341$, $n = 26$, $p < .00005$). Errors on the problems with no valid answers were mainly consistent with the premises rather than inconsistent with them: 76% consistent errors in Experiment 1 (Wilcoxon's $T = 269$, $n = 27$, $p < .03$); 99% consistent errors in Experiment 2 ($p = .533$); and 85% consistent errors in Experiment 3 (Wilcoxon's $T = 103$, $n = 14$, $p < .0003$).

The general pattern of the results supports the conclusion that one-model problems lead to fewer errors than multiple-model problems. Indeed, the number of models seems to be a more critical factor than the size of the models (3 events vs. 5 events), because there was no significant difference between the two-premise and four-premise problems. The results also suggest that the subjects generally interpreted the premises in the order in which they were stated: if they had worked backwards from the question, then they could have ignored the irrelevant premise in the multiple-model problems with valid answers and thereby reduced them to one-model problems interrelating only four events.

Is it possible that some difference in the problems, other than the number of models, yields an alternative explanation of our results? The problems were matched for the relational terms they contained. Thus, for example, there was a problem based on two premises with "before" for each of the three sorts of problem, and a problem with "before" in the first premise and "after" in the second premise, and so on for all four possible combinations of "before" and "after". The relational term in the remaining premises was always "while". The only difference between one-model problems and multiple-model problems with valid answers was in the disposition of events, for example, the former had the second premise:

b happens before c

the latter had the converse:

c happens before b

Hence, the results cannot be explained by differences in either lexical marking or the congruence between temporal order and order of mention of events in premises. Although the overall data favored the model theory, they certainly do not eliminate theories based on formal rules (e.g., Braine et al., 1984; Rips, 1994). We accordingly tried to make a direct comparison between the two sorts of theory in the next experiment.

EXPERIMENT 4

The aim of the experiment was to contrast the predictions based on the lengths of formal rule derivations with those based on the mental model theory. Consider the following "transitive" sort of one-model problem, which had not been used in the previous experiments:

a happens before b
b happens before c
d happens while a
e happens while c
What is the relation between d and e?

The formal derivation of the answer must first establish the relation between a and c, from the transitivity of "a before b" and "b before c", and then use this derived relation "a before c" to establish the relation between d and e. Now consider the following multiple-model problem of the same form as problem 2 used in the previous experiments, though we have relabeled the events in order to clarify a point about its formal derivation:

b happens before c
a happens before c
d happens while a
e happens while c
What is the relation between d and e?

The problem has a valid answer and its formal derivation is just part of the derivation for the previous problem, because there is no need to derive a transitive relation: the key relation between a and c is explicitly asserted by the second premise. In other words, for the one-model problem, reasoners need to prove two relations – first, the relation between a and c, and then, using this information, the relation between d and e. For the multiple-model problem, however, reasoners are given the relation between a and c in a premise, and so they need to prove only the relation between d and e. Hence, the multiple-model problem should be easier because its derivation is included within the derivation for the one-model problem. The model theory, of course, makes the opposite prediction. The first problem has one model, whereas the second problem yields at least two alternative models:

b	a	c	a	b	c
	d	e	d		e

which both support the answer:

d happens before e

The predictions based on formal derivations and on mental models are accordingly diametrically opposed to one another: formal derivations predict that one-model problems should yield *more* errors than multiple-model problems, whereas mental models predict that one-model problems should yield *fewer* errors than multiple model problems. The experiment tested these contrasting predictions. It also included multiple-model problems with no valid answers, which both theories predict should yield the greatest number of errors.

5. Method

5.1. Design

The subjects acted as their own controls and carried out eight versions of each of three sorts of deduction: "transitive" one-model problems with valid answers, multiple-model problems with valid answers, and multiple-model problems with no valid answers (see Table 1 for a summary of these types of problem). The order of presentation of the 24 problems was randomized for each subject.

5.2. Materials

The lexical materials concerned the temporal order of daily activities in the morning by one or two persons as in Experiment 2, and the eight different versions of each sort of problem were constructed in the same way as in Experiment 3. The main clause was prior to the subordinate clause in each premise.

5.3. Procedure

The procedure was the same as in the previous experiments, except that the subjects were tested in four small groups of four subjects in order to ensure that the subjects completed all the problems.

5.4. Subjects

Sixteen subjects carried out the experiment. They were all university students without a training in logic. They were paid five dollars per hour to participate in the experiment, which lasted for about 35 minutes.

6. Results and discussion

As Table 2 shows, the subjects solved 94% of the one-model problems, 84% of the multiple-model problems with valid answers, and 19% of the multiple-model problems with no valid answers. This trend was reliable by subjects (Page's $L = 217$, $n = 16$, $p < .001$) and by materials (Page's $L = 110.5$, $n = 8$, $p < .001$).

The one-model problems yielded fewer errors than the multiple-model problems with valid answers (Wilcoxon's $T = 56$, $n = 11$, $p < .03$, with only two subjects violating the prediction). Eighty-two percent of the errors to the problems with no valid answers were consistent with the premises, and only 18% of the errors were inconsistent with the premises (Wilcoxon's $T = 98$, $n = 15$, $p < .02$).

The pattern of results clearly corroborates the predictions of the model theory and runs counter to those based on formal derivations. Transitive one-model problems are answered correctly more often than multiple-model problems, which are answered correctly more often than problems without a valid answer. One unexpected result was the very poor performance with multiple-model problems with no valid answers. One possibility is that the subjects were biased against responding that there was no valid answer to a problem – a bias that might have been elicited by the fact that the practise problems had valid answers (Braine, personal communication). Both the model theory and formal rule theories predict that these problems should be hardest. The subjects tended to give answers to these problems that were consistent with the premises rather than inconsistent with them, as predicted by the model theory. They evidently based their answers on one model of the premises and overlooked the others. Since the instructions and materials were identical to those of the previous experiment, it seems that the large proportion of errors on these problems is a chance fluctuation. The fact that the subjects were drawn from a variety of university disciplines does not seem to be critical (cf. the results of the next experiment).

Although the results cast doubt on the predictive value of formal derivations, they do not count against formal rule theories per se. There is at least one alternative explanation. The multiple-model problems with valid answers had one irrelevant premise, that is, the first premise, whereas the one-model problems had no irrelevant premises. The presence of an irrelevant premise might confuse subjects, and make it harder to find a derivation of the answer (Rips, 1994). Hence, it could account for the difference that we observed between the two sorts of problem. Our final experiment was designed both to test this explanation and to make a more stringent test of the competing predictions.

EXPERIMENT 5

The main aim of this experiment was to assess how long subjects take to read the premises of temporal deductions and to respond with their answers. If the model theory is correct, the time taken to read a premise that calls for the construction of multiple models should be longer than the time taken to read a premise that calls for the construction of one model. This prediction has never been examined before, but it is a strong test of the model theory because it should take longer to set up alternative models.

We examined four sorts of deductions:

1. One-model problems with an irrelevant premise as in Experiments 1–3, which do not call for a transitive inference:

a happens before b
b happens before c
d happens while b
e happens while c
What is the relation between d and e?
This problem has the following model:

a b c

d e

2. Transitive one-model problems with no irrelevant premise, as in the previous Experiment:
a happens before b
b happens before c
d happens while a
e happens while c
What is the relation between d and e?
This problem has the following model:

a b c

d e

3. Multiple-model problems with a valid answer, which as in all the previous experiments contain an irrelevant premise:
a happens before c
b happens before c
d happens while b
e happens while c
What is the relation between d and e?
This problem has at least the following two models:

a b c b a c

d e d e

which support the answer: d happens before e. The second premise in problems of this sort calls for the construction of two alternative models, because of the indeterminacy between a and b. Hence, it should take longer to read than the second premise of the one-model problems.

4. Multiple-model problems with no valid answer, as in the previous experiments:
a happens before c
b happens before c
d happens while b

e happens while a
What is the relation between d and e?
This problem has the following alternative models:

a b c b a c

e d d e

which do not support a valid answer to the question. The second premise again calls for the construction of alternative models.

The model theory predicts that the one-model problems, whether transitive or non-transitive, should be easier than the multiple-model problems. If formal derivations are the critical factor, however, then the task of drawing the transitive inference to interrelate a and c adds extra steps to the derivation, and so transitive one-model problems should be harder than non-transitive problems, whether one-model or multiple-model. The two accounts therefore make distinct partial rank-order predictions. The model theory predicts the following order in the proportions of correct answers to the four sorts of problem: 1, $2 > 3 > 4$; whereas formal derivations predict the following rank order: 1, $3 > 2 > 4$, granted that the problems with no valid answers should yield the most errors because one has to search the space of possible derivations exhaustively. If the key factor is the presence of an irrelevant premise then transitive one-model problems (2), which lack such a premise, should be easier than one-model problems (1) and multiple-model problems with a valid answer (3). The problems with no valid answers (4) ought also to be easy, but should be excluded from this comparison because of the different sort of correct response. The experiment examined these predictions for both the subjects' latencies of response and their accuracy in responding.

7. Method

7.1. Design

The subjects acted as their own controls and carried out eight versions of each of four sorts of deduction: transitive one-model problems, non-transitive one-model problems, multiple-model problems with valid answers, and multiple-model problems with no valid answers (see Table 1).The order of presentation of the 32 problems was randomized for each subject.

7.2. Materials

The lexical materials concerned the temporal order of daily activities by two persons, and the eight different versions of each sort of problem were constructed

in the same way as in Experiments 3 and 4. The main clause was prior to the subordinate clause in each premise.

7.3. Procedure

The subjects were tested individually, and the experiment was carried out using a computer. At the beginning of each trial, the screen signaled "press space-bar for the next problem". When subjects pressed the space-bar, the first premise appeared and stayed on the screen until the subjects pressed the space-bar again. At this point, the next premise appeared, and the procedure continued until the fourth premise. When the subjects then pressed the space-bar, the question appeared and stayed on the screen until they pressed the space-bar again. The subjects then typed their answer. The computer recorded five main latencies: the time taken to read each of the four premises, and the time from the presentation of the question until the subjects began to type their response. (The computer also recorded the time that the subjects took to type their complete answer, but we did not use this measure because the length of the responses differed between problems with valid answers and those with no valid answers.) The instructions, which were presented on the VDU, were the same as in the previous experiment except that the subjects were told that although they would be timed, they should concentrate on making the correct responses. The instructions also explained how to use the space-bar to present the next premise and how to respond. The subjects solved one practise problem to make sure that they had grasped the procedure. None of the subjects had any difficulty with it.

7.4. Subjects

Twenty-four subjects carried out the experiment. They were all university students without a training in logic. None of them participated in the previous experiments. They were paid about five dollars per hour to participate in the experiment, which lasted about 40 minutes.

8. Results and discussion

Overall, as Table 2 shows, the subjects solved 93% of the transitive one-model problems, 89% of the non-transitive one-model problems, 81% of the multiple-model problems with valid answers, and 44% of the multiple-model problems with no valid answers. There was no reliable difference in accuracy of solutions between the two sorts of one-model problem (Wilcoxon's $T = 50$, $n = 11$, n.s.). However, there was the following reliable trend in correct answers: one-model problems were easier than multiple-model problems with valid answers, which were easier than multiple-model problems with no valid answers (by subjects, Page's $L = 315$, $n = 24$, $p < .0001$; and by materials, Page's $L = 111$, $n = 8$, $p < .001$). In addition, the one-model problems were answered correctly more

often than multiple-model problems with valid answers (Wilcoxon's $T = 167$, $n = 20$, $p < .02$). The pattern of correct answers accordingly corroborated the model theory and ran counter to the predictions based formal derivations. Eighty-one percent of the erroneous answers to problems with no valid answers were consistent with the premises and only 19% were inconsistent with them, though the two sorts of error are equiprobable a priori (Wilcoxon's $T = 152$, $n = 19$, $p < .02$).

Fig. 2 presents the means of the reading times for the four premises for the one-model problems, the multiple-model problems with valid answers, and the multiple-model problems with no valid answers. For legibility, the figure is a graph rather than a histogram, and likewise because there was no reliable difference at any point between the two sorts of one-model problem, we have collapsed the data. (The mean reading times in seconds for the one-model problems were as follows with those for non-transitive problems with an irrelevant

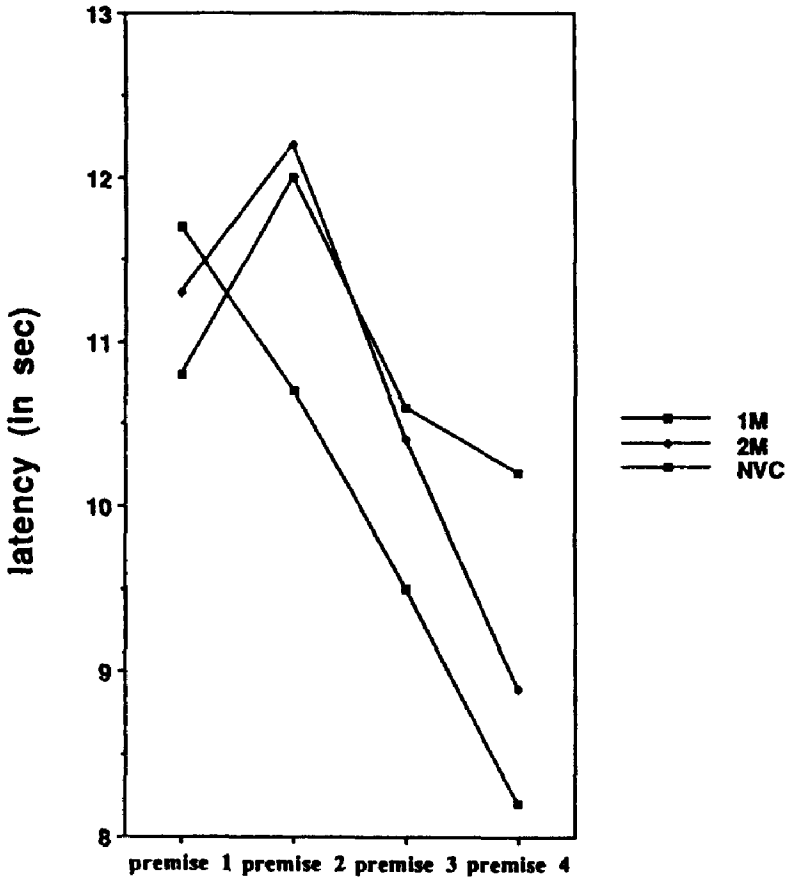

Fig. 2. The mean reading times in Experiment 5 for each of the four premises for the one-model problems (1M), the multiple-model problems with valid answers (2M), and the multiple-model problems with no valid conclusions (NVC).

premise preceding those for the transitive problems: first premise 12.0 and 11.4, second premise 10.9 and 10.5, third premise 9.5 and 9.5, fourth premise 8.6 and 7.8.) The mean latencies to respond to the questions for the four sorts of problems were as follows:

Non-transitive one-model problems: 5.8 seconds
Transitive one-model problems: 7.0 seconds
Multiple-model problems with valid answers: 8.7 seconds
Multiple-model problems with no valid answers: 10.7 seconds

We carried out two analyses of variance (using the Statistica general MANOVA program): the first analysis was of the reading times for each of the four premises, and the second analysis was of the latencies to respond to the question. The analysis of the reading times showed that there was a significant effect of the type of problem ($F(3, 69) = 3.82$, $p < .02$), a significant difference over the four premises; i.e., the subjects read them progressively faster ($F(3, 69) = 5.68$, $p < .002$), and a significant interaction between these two variables ($F(9, 207) = 2.08$, $p < .04$). The analysis of the response times to the questions showed that there was a significant effect of the type of problem ($F(3, 69) = 7.89$, $p < .0002$). We carried out three planned orthogonal comparisons on the reading times for each premise and on the latencies to respond to the question:

1. The difference between the two sorts of one-model problem. As we expected, it was not significant in any case, though it was nearly so for the response time to the question ($F(1, 23) = 3.05$, $p < .1$). Subjects accordingly showed a slight tendency to be faster with the problems with an irrelevant premise, but to be less accurate with them. There may be a trade-off between speed and accuracy here, but we emphasize that neither difference was reliable.
2. The difference between one-model problems and multiple-model problems with valid answers. The one-model problems had shorter latencies than the multiple-model problems for the reading times of the second premise ($F(1, 23) = 6.3$, $p < .02$) and for the response times to the question ($F(1, 23) = 7.03$, $p < .02$), but not in any other case. The difference in the reading times of the second premise corroborates the crucial prediction. It is this premise that calls for the construction of alternative models in the multiple-model case, but not in the one-model case.
3. The difference between the three problems with valid answers and the multiple-model problem with no valid answer. As expected, it was significant for the response time to the question ($F(1, 23) = 9.52$, $p < .006$).

Previous results in studies of temporal connectives have suggested that "before" is the lexically unmarked term and therefore slightly easier to understand than the marked term "after" (see, for example, Clark, 1971). We therefore examined the difference between the reading times for "before" and "after" (in the first premise so that the measure would not be influenced by the stage or sort

of model-building operation). The mean reading time for premises with "before" was 10.7 seconds and the mean reading time for premises with "after" was 12.1 seconds (Wilcoxon's $T = 268$, $n = 24$, $p < .0003$). The main clause always preceded the subordinate clause, and so this result could be because "after" is the marked term and more difficult to understand, or because the order of mention of the two events in the "after" premise is opposite to their actual temporal sequence (see Clark and Clark, 1968; Smith and McMahon, 1970; Mandler, 1986). One further piece of evidence in favor of the relative ease of coping with "before" is that the subjects showed a bias in their answers towards using "before" (55% of responses) rather than "after" (14% of responses, Wilcoxon's $T = 275$, $n = 23$, $p < .00003$). One further piece of evidence in favor of the congruence between order of mention and temporal order concerns the two sorts of question. For the one-model problems, subjects showed a tendency to respond faster to those questions in which the two events were mentioned in the same order as their temporal occurrence (5.9 seconds) than to those questions in which the two events were mentioned in the opposite order to their temporal occurrence (6.6 seconds, but the difference was not quite significant, Wilcoxon's $T = 200$, $n = 24$, $p > .07$). Nevertheless, both factors may affect the difficulty of interpreting individual premises.

GENERAL DISCUSSION

The experimental results establish three main phenomena, and they corroborate the hypothesis that reasoning about temporal relations depends on mental models of the sequences of events. The first phenomenon concerns the number of models. When a description is consistent with just one model, the reasoning task is simple and subjects typically draw over 90% correct answers. When a description is consistent with more than one model, there is a reliable decline in performance. Experiments 4 and 5 pitted the predictions of the model theory against contrasting predictions based on formal derivations. The results showed that the one-model problems were reliably easier than the multiple-model problems, even though the one-model problems call for formal derivations that add extra steps to the derivations for the multiple-model problems. For multiple-model problems that have a valid answer, the materials in all the experiments have the following property: if reasoners construct just one of the possible models of the premises and base their answer on this model, the answer will still be correct. Hence, unlike multiple-model syllogisms or propositional deductions, it is not necessary to construct each model in the multiple-model case in order to reach the correct answer. Nevertheless, our results suggest that subjects attempted to do so, and indeed they have no way of knowing that one model will suffice. Because a failure to consider all models can still yield the correct answer, the model theory predicts that reasoning in this case should not yield vastly more errors than the one-model case. In contrast, it is vital to consider the multiple models for those problems that have no valid answer, and they generally yielded many more errors, especially in Experiment 4.

The second phenomenon concerns the subjects' erroneous answers. Current versions of formal rule theories make no specific predictions about the nature of such answers (Evans, 1991): subjects are said to err because they misapply a rule or fail to find a correct derivation. The model theory, however, predicts that erroneous answers arise because reasoners fail to consider all the models of the premises, and so these answers should tend to be consistent with at least one model of the premises rather than inconsistent with all of them. The results corroborated this prediction of the model theory. However, if current formal rule theories were modified to incorporate a "censor" that checked for contradictions between conclusions and premises, then, as Martin Braine argues (personal communication), they too would make the same prediction.

The third phenomenon concerns the time that subjects took to read the premises and to respond to the questions (in Experiment 5). In general, the subjects read the four premises progressively faster, but contrary to this trend they took reliably longer to read a premise that led to multiple models than to read a corresponding premise in a one-model problem. Formal rule theories make no such prediction, and it is hard to reconcile this result with such theories because they make no use of models. The result also suggests that subjects do not construct models that represent indeterminacies directly within a model (akin to Fig. 1). Otherwise, the subjects would not have taken longer to read the second premise of multiple-model premises, they would not have been more prone to err with them, and they would not have taken longer to answer questions about them.

Our experiments were based on the assumption that a model of five events was small enough to be accommodated within working memory. Obviously, a one-model problem based on, say, a hundred premises would not be, and so would be very difficult. Hence, our claim is that provided models are tractable in this way, the key factor is the number of models to be constructed. Skeptics might argue that even in Experiments 4 and 5 a process of formal reasoning might be necessary to establish transitivity in the case of the transitive one-model problems. Our failure to find any reliable difference between the transitive one-model premises and the non-transitive one-model problems with irrelevant premises counts against this hypothesis. Are there any differences between one-model and multiple-model problems that could provide the basis for an alternative explanation of our results? In discussing the results of Experiments 1–3, we showed that the lexical materials and the congruence of order of mention and temporal order could not account for the phenomena. Likewise, violations of the optimal referential structure in which to understand discourse (see, for example, the "given-new" contract of Clark and Haviland, 1977) cannot account for the difference, because the referential structures are the same for the first two premises of both the one-model and the multiple-model problems. In Experiment 5, for example, there are one-model problems with the following two initial premises:

a happens before b
c happens after b

which have the same referential structure as the premises:

a happens before b
c happens before b

for the multiple-model problems, and yet, as Fig. 2 shows, the second premise of the multiple-model problems takes longer to read than the second premise of the one-model problems. Another potential factor is the presence of an irrelevant premise. It might lead reasoners up the "garden path" and thus make it harder for them to find a correct derivation. This hypothesis could account for the difficulty of the multiple-model problems with valid answers in Experiment 4. It runs into difficulties, however, with our other findings. In Experiments 1–3, there was an irrelevant premise in both the one-model and the multiple-model problems with valid answers, and yet the one-model problems were easier than the multiple-model problems. Likewise, in Experiment 5, there were one-model problems with, and without, irrelevant premises, and there was no reliable difference between them. We conclude that an irrelevant premise may have a marginal effect on reasoning, but it cannot explain the difference between one-model and multiple-model problems.

In a study that complements our own, Vandierendonck and De Vooght (1992) observed that one-model problems led to fewer errors than multiple-model problems, and they reported a finding relevant to the use of time in temporal representations: transitive one-model problems tended to be as hard as multiple-model problems with valid conclusions. This result might be expected if the subjects were imagining the events one after another in an actual *kinematic* sequence: a b c, because in both cases it is necessary for the whole sequence to unfold. One note of caution, however, is that their experimental procedure seems to have reduced performance considerably on all one-model problems (only about 60% correct choices of conclusion in most of the experiments). It may be that different procedures will increase or decrease the propensity of individuals to imagine events one after another rather than to use static representations of temporal relations.

One other factor is likely to affect the difficulty of temporal reasoning, namely, the referential coherence of the premises. In our studies, the problems were referentially coherent, that is, in premises such as:

a happens before b
b happens before c
d happens while a
e happens while c

each premise (apart from the first one) refers back to an event in an earlier premise. Other studies have shown that when the assertions in a spatial description are not presented in a referentially coherent order, the passage takes longer to read and is harder to understand (Ehrlich and Johnson-Laird, 1982). Likewise, Vandierendonck and De Vooght (1992) observed that disruptions to referential coherence produced still more errors in their reasoning task.

In certain reasoning problems, it is advantageous to ignore irrelevant premises and to deal with the relevant premises in a referentially coherent order. Our computer model of temporal reasoning resorts to this strategy whenever its normal mode of processing leads to more models than its index of working memory capacity. The effect of this strategy in Experiments 1–4 would have been to reduce multiple-model problems with irrelevant premises to one-model problems. The experimental data accordingly imply that the subjects did not in general resort to such a strategy. The presentation of the premises one at a time in Experiment 5 render the strategy still less likely.

In conclusion, temporal reasoning can be explained by the theory of mental models: reasoners appear to construct static mental models akin to those used for spatial reasoning. Our computer program reasons in this way, and the results of the first experiments to examine temporal reasoning bore out three predictions of the theory. First, multiple models mean more work: subjects take longer and are more likely to err. Second, erroneous answers tend to be consistent with some models of the premises rather than inconsistent with all of them. Third, an assertion that calls for the construction of multiple models takes longer to read than does a comparable assertion in a one-model problem. Formal rules do not make these predictions, but some formal rule theorists allow that mental models can play a part in deductive reasoning, and Braine (1994)p. 245) himself writes: "I have little doubt that much reasoning does use mental models".

Acknowledgments

We thank Andreas de Troy for implementing the program used to carry out Experiment 5 and Malcolm Bauer for help with the statistical analyses. We thank Jonathan Baron, Martin Braine, Jean Mandler, Jacques Mehler, Victoria Shaw, and Steven Sloman, for their helpful comments on earlier drafts. Schaeken is supported by the National Fund for Scientific Research of Belgium, and Johnson-Laird's research is supported in part by the John S. McDonnell foundation.

References

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the Association for Computing Machinery, 26,* 832–843.

Braine, M.D.S. (1994). Mental logic and how to discover it. In J. Macnamara and G.E. Reyes (Eds.), *The logical foundations of cognition* (pp. 241–263). Oxford: Oxford University Press.

Braine, M.D.S., Reiser, B.J., & Rumain, B. (1984). *Some empirical justification for a theory of natural propositional logic. The psychology of learning and motivation* (Vol. 18, pp. 313–371). New York: Academic Press.

Byrne, R.M.J., & Johnson-Laird, P.N. (1989). Spatial reasoning. *Journal of Memory and Language, 28,* 564–575.

Clark, E.V. (1971). On the acquisition of the meaning of "before" and "after". *Journal of Verbal Learning and Verbal Behavior, 10,* 266–275.

Clark, H.H., & Clark, E.V. (1968). Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology, 20,* 129–138.

Clark, H.H., & Haviland, S.E. (1977). Comprehension and the given-new contract. In R.O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Norwood, NJ: Ablex.

Dinsmore, J. (1991). *Partitioned representations.* Dordrecht: Kluwer.

Ehrlich, K., & Johnson-Laird, P.N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior, 21,* 296–306.

Evans, J.St.B.T. (1991). Theories of reasoning: The fragmented state of the art. *Theory and Psychology, 1,* 83–105.

Fraisse, P. (1963). *The psychology of time.* New York: Harper and Row.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78,* 111–133.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), *Studies in syntax. Vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hagert, G. (1984). Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In T. O'Shea (Ed.), *Advances in artificial intelligence* (pp. 389–398). Amsterdam: North-Holland.

Huttenlocher, J. (1968). Constructing spatial images: a strategy in reasoning. *Psychological Review, 75,* 286–298.

Isard, S.D. (1974). What would you have done if...? *Theoretical Linguistics, 1,* 233–255.

Isard, S.D., & Longuet-Higgins, H.C. (1971). Modal tic-tac-toe. In R.J. Bogdan & I. Niiniluoto (Eds.), *Logic, language, and probability* (pp. 287–296). Dordrecht: Reidel.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness.* Cambridge, MA: Harvard University Press/Cambridge, UK: Cambridge University Press.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N., Byrne, R.M.J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99,* 418–439.

Johnson-Laird, P.N., Byrne, R.M.J., & Tabossi, P. (1989). Reasoning by model: the case of multiple quantification. *Psychological Review, 96,* 658–673.

Johnson-Laird, P.N., & Savary, F. (1995). How to make the impossible seem probable. In J.D. Moore and J.F. Lehman (Eds.), *Proceedings of the annual conference of the cognitive science society* (pp. 381–390). Pittsburgh, PA.

Macnamara, J. (1986). *A border dispute: The place of logic in psychology.* Cambridge, MA: Bradford Books, MIT Press.

Mandler, J.M. (1986). On the comprehension of temporal order. *Language and Cognitive Processes, 1,* 309–320.

Miller, G.A., & Johnson-Laird, P.N. (1976). *Language and perception.* Cambridge, UK: Cambridge University Press/Cambridge, MA: Harvard University Press.

Newstead, S.E., Manktelow, K.I., & Evans, J.St.B.T. (1982). The role of imagery in the representation of linear orderings. *Current Psychological Research, 2,* 21–32.

Oakhill, J., & Garnham, A. (1985). Referential continuity, transitivity, and the retention of relational descriptions. *Language and Cognitive Processes, 1,* 149–162.

Ohlsson, S. (1984). Induced strategy shifts in spatial reasoning. *Acta Psychologica, 57,* 46–67.

Osherson, D. (1975). Logic and models of logical thinking. In R.J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 81–91). Hillsdale, NJ: Erlbaum.

Page, E.B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association, 58,* 216–230.

Piaget, J. (1969). *The child's conception of time.* London: Routledge and Kegan Paul. (Originally published 1927.)

Pollock, J. (1989). *How to build a person: A prolegomenon.* Cambridge, MA: MIT/Bradford Books.

Reichenbach, H. (1947). *Elements of symbolic logic.* New York: Free Press.

Richardson, J.T.E. (1987). The role of mental imagery in models of transitive inference. *British Journal of Psychology,* 78: 189–203.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90,* 38–71.

Rips, L.J. (1994). *The psychology of proof.* Cambridge, MA: Bradford Books/MIT Press.

Smith, K.H., & McMahon, L.E. (1970). Understanding order information in sentences: Some recent work at Bell Laboratories. In G.B. Flores d'Arcais and W.J.M. Levelt (Eds.), *Advances in psycholinguistics* (pp. 253–274). Amsterdam: North-Holland.

Steedman, M.J. (1982). Reference to past time. In R.J. Jarvella & W. Klein (Eds.), *Speech, place, and action* (pp. 125–157). London: Wiley.

Vandierendonck, A., & De Vooght, G. (1992). *Is reasoning with time concepts based on a spatialized representation of time?* Report of the Department of General Psychology, University of Ghent, Belgium.