COGNITIVE SCIENCE Vol 23 (3) 1999, pp. 247–303 ISSN 0364-0213

Strategies in Syllogistic Reasoning

MONICA BUCCIARELLI
Centro di Scienza Cognitiva Universita' di Torino

P. N. JOHNSON-LAIRD
Princeton University

This paper is about syllogistic reasoning, i.e., reasoning from such pairs of premises as, All the chefs are musicians; some of the musicians are painters. We present a computer model that implements the latest account of syllogisms, which is based on the theory of mental models. We also report four experiments that were designed to test this account. Experiments 1 and 2 examined the strategies revealed by the participants' use of paper and pencil as aids to reasoning. Experiment 3 used a new technique to externalize thinking. The participants had to refute, if possible, putative conclusions by constructing external models that were examples of the premises but counterexamples of the conclusions. Experiment 4 used the same techniques to examine the participants' strategies as they drew their own conclusions from syllogistic premises. The results of the experiments showed that individuals not trained in logic can construct counterexamples, that they use similar operations to those implemented in the computer model, but that they rely on a much greater variety of interpretations of premises and of search strategies than the computer model does. We re-evaluates current theories of syllogistic reasoning in the light of these results.

I. INTRODUCTION

The more psychologists study certain topics, the less they seem to know about them. Syllogisms may be a case in point. Syllogisms are a small set of inferences that are based on two premises, each containing a single quantifier, e.g.,

Some actuaries are businessmen.
All businessmen are conformists.
∴ Some actuaries are conformists.

Direct all correspondence to: P. N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, NJ 08544; E-Mail: phil@clarity.princeton.edu; Monica Bucciarelli, Centro di Scienza Cognitiva, Universitadi Torino.

Syllogisms were first analyzed by Aristotle, and traditionally their premises and conclusions are in one of four moods:

All X are Y	(abbreviated A)
Some X are Y	(abbreviated I)
No X are Y	(abbreviated E)
Some X are not Y	(abbreviated O)

For a valid deduction, the two premises must contain at least one term in common—the so-called middle term (designated B) and two end terms (designated A and C) that each occur in a single premise. The three terms in the premises can be in one of four possible arrangements, or figures:

1.	2.	3.	4.
A–B	B–A	A–B	B–A
B–C	C–B	C–B	B–C

We will use this numbering system in the present paper because it demonstrates that reasoners may draw either of the two sorts of conclusion: A–C and C–A. Thus, the example above is in Figure 1. However, the scholastic logicians considered the figure of a syllogism to include the conclusion, and they used a different numbering system than the one above. Figure affects syllogistic reasoning (see e.g., Johnson-Laird & Bara, 1984), but the cause of the effects is controversial. They also occur with three-term series problems, with various sorts of multiply quantified premises, and with conditionals and disjunctions. Because the effects are not unique to syllogisms, we have chosen to focus on other aspects of syllogistic reasoning in the remainder of the paper.

When you first encountered a syllogism, such as our opening example, you might have had the following thought. It is obviously valid—indeed, it looks trivial from a psychological point of view. It also seems remote from the sort of reasoning that goes on in daily life. You might suppose that trivial inferences remote from daily life are not worth studying. But, a closer examination shows that you would be wrong to conclude that syllogisms are not worth studying. Even though they are logically transparent, they are far from psychologically trivial. Psychologists have been studying syllogisms for nearly a hundred years (see e.g., Störring, 1908), but have yet to reach a consensus about how individuals not trained in logic cope with them. Likewise, syllogisms are not so remote from the inferences of daily life. They seem more regimented than everyday reasoning, but that is a result of their surface form. The underlying logical relations occur all the time. For example, we can re-express the preceding inference in a syllogistic guise:

All syllogisms are trivial inferences remote from daily life.
No trivial inferences remote from daily life are worth studying.
∴ No syllogisms are worth studying.

Early psychological studies of syllogisms were concerned with the factors that lead reasoners astray, the atmosphere of the premises (e.g., Woodworth & Sells, 1935), the illicit conversion of premises (e.g., Chapman & Chapman, 1959), and the adverse effects of beliefs and prejudices (e.g., Henle & Michael, 1956). Twenty years ago, however, psychologists proposed the first theories of how reasoners might reach valid conclusions (e.g., Erickson, 1974; Johnson-Laird, 1975; Revlis, 1975). Since then, there has been a

plethora of theories, some based on Euler circles (Fisher, 1981; Ford, 1995; Stenning & Oberlander, 1995); some based on Venn diagrams (Guyote & Sternberg, 1981; Newell, 1981); some based on mental models (Cardaci, Gangemi, Pendolino, & Di Nuovo, 1996; Johnson-Laird & Bara, 1984; Polk & Newell, 1995); some based on formal rules of inference (Braine & Rumain, 1983; Rips, 1994); and some based on the idea that individuals are not reasoning, but are following the atmosphere of the premises (Wetherick & Gilhooly, 1990; Martin Levine, personal communication, 1994) or selecting a conclusion that matches the mood of the least informative premise (Chater & Oaksford, 1999). The variety of theories confirms that, even though syllogisms are logically simple, they are psychologically complex.

In the present paper, we aim to make progress in resolving the theoretical controversy. We begin with a computer program implementing the mental model theory of syllogistic reasoning, and then report four experiments designed to test whether this program gives an accurate account of the strategies, representations, and procedures that individuals not trained in logic use in syllogistic reasoning. In the first two experiments, the participants were allowed to use paper and pencil as they reasoned. The results corroborated certain aspects of the model theory, but they also supported some rival accounts. We describe a new technique designed to externalize the process of thought in syllogistic reasoning, and report a third experiment that used this technique to investigate whether people can refute putative conclusions. This study was motivated in part by Polk's (1993) claim that the search for counterexamples appeared to underlie few predictions and that his own model theory provided a better account of individual differences when it dropped this component (see also Polk & Newell, 1995). Martín-Cordero and Gonzàlez-Labra (1994) argued similarly that a major flaw in the mental model theory is its failure to specify how human reasoners search for counterexamples. The problem is not a conceptual one because the computer program contains a well-specified search procedure. The difficulty is instead in obtaining evidence about the process of searching for alternative models of premises. Our fourth experiment, therefore, used the externalization technique to examine how the participants drew their own syllogistic conclusions. The experiments showed that reasoners do construct multiple models of the premises, but use a striking variety of strategies. We conclude with a re-evaluation of all the current theories of syllogistic reasoning.

II. A COMPUTER IMPLEMENTATION OF THE MENTAL MODEL THEORY

The fundamental idea underlying the mental model theory is that people interpret assertions by constructing models of the corresponding situations (Johnson-Laird, 1983). Thus, given the assertion:

All the artists are beekeepers

reasoners imagine a situation in which there is a small, but arbitrary, number of artists and beekeepers, and in which each of the artists is a beekeeper. The form of this representation is problematic. Indeed, the problems in developing a model theory of syllogisms can be illustrated by considering the possible representations of assertions of the form:

All A are B

given the existence of As and Bs. An obvious representation is to use Euler circles (see e.g., Erickson, 1974). One problem with them, as Erickson (1974) showed, is that they can lead to a combinatorial explosion (but for alternative algorithms that obviate this problem, cf. Ford, 1995; Stenning & Oberlander, 1995). Another problem, as Rips (1994) has argued, is that reasoners not trained in logic are unlikely to use Euler circles unless they have been taught the technique: it took a mathematical genius, Leibniz, to invent them. The major disadvantage of Euler circles, however, is that they do not generalize to relational inferences, such as the following example (see Russell, 1946):

All horses are animals.

∴ All horses' heads are animals' heads.

Hence, the model theory postulates that finite sets of entities are represented by finite sets of mental tokens that readily accommodate relations among entities (see Johnson-Laird, 1983).

Models could be based on the principle that each set of entities is represented in its entirety. Thus, a model of, All A are B, could take the form:

a b
a b
a b

where each row represents a separate individual, and the model is based on the arbitrary assumption that there are three As in the situation. It is possible that there are Bs that are not As, and so this possibility would have to be represented in a separate model:

a b
a b
a b
b
b

By assumption, the set of As is exhaustively represented, and so the new tokens of Bs could be fleshed out explicitly to represent that they are not As:

a b
a b
a b
−a b
−a b

where '−' denotes negation. One problem with these sorts of models is that they are isomorphic to Euler circles and, therefore, run the risk of a combinatorial explosion. They are also psychologically implausible because the numbers of models required for different syllogisms do not correspond to their psychological difficulty. Accordingly, another possibility is to represent that some entities are optional, i.e., they may or may not be in the situation:

a b
a b
a b
(b)
(b)

where the items in parentheses denote optional entities (for variants of this assumption, see Johnson-Laird, 1975; Johnson-Laird and Bara, 1984; Johnson-Laird & Steedman, 1978). But, models containing optional entities also run afoul of the psychological results. In particular, the preceding model suggests that reasoners should not confuse All As are Bs, with its converse, All Bs are As. In fact, reasoners often make such erroneous inferences.

In extending the model theory to reasoning based on sentential connectives, Johnson-Laird and Byrne (1991) proposed that conditionals of the form:

If A then B

are represented by an explicit model of the case where the antecedent is true:

a b

and an implicit model of the case, or cases, in which the antecedent is false:

. . .

where the ellipsis denotes a model that has no explicit content. Reasoners must make a mental footnote that the explicit model exhausts the cases where the antecedent, A, is true. The footnote is important if reasoners need to flesh out explicitly the possibilities in which the antecedent is false, and square brackets represent this footnote:

[a] b
. . .

There is a close relation between conditionals, such as:

If it is a dog, then it is a mammal.

and quantified assertions, such as:

All dogs are mammals.

Hence, Johnson-Laird and Byrne (1991) adopted mental footnotes for syllogisms, and they proposed that All A are B is represented by the following sort of model:

[a] b
[a] b

in which there are two As. The mental footnote denoted by the square brackets denotes that the set of As has been represented in its entirety. Such footnotes, however they are actually represented in the mind, are more readily forgotten than explicit tokens. The tendency to forget them leads to systematic fallacies in reasoning with quantified assertions (see Yang & Johnson-Laird, unpublished data). Their ephemeral nature is also reflected in the following computer program, particularly in the procedure for drawing conclusions from models.

The computer program works according to the latest version of the model theory (see Bara, Bucciarelli, and Johnson-Laird, 1995). Because our results will demonstrate the inadequacies of the program, we will try to convey only its general principles and their motivation. We assume throughout that there is no doubt about the existence of members of the sets referred to in the premises. Given a premise of the form:

All A are B

the program constructs a small number of As, ensures that each of them is a B, and represents that the set of As has been exhaustively represented:

[a] b
[a] b

where each line denotes a separate entity in the model, and the square brackets indicate that the set of As has been represented in its entirety. Hence, no new tokens of As can be added to the model, i.e., if new entities are introduced into the model, they cannot be As. A premise of the form:

Some A are B

could have the form:

a b

But, in our view, people are likely to envisage explicitly that there are As that are not Bs, and Bs that are not As, and so the program constructs the following sort of initial model:

a b
a
b

Each entity in a model represents what is necessary given the premise, but the present model can have additional tokens added to it to create either of the following alternative models:

a b a b
a b a −b
a b −a b

Hence, the only type of entity that is necessary is: a b. And, as in logic, "some" is treated as equivalent to "at least some", which is compatible with "all." Although the actual numbers of tokens are, in theory, arbitrary, we simplify the operations of the program without affecting its outcomes by using the same number of initial tokens for each of the three terms in a syllogism, and so we chose two tokens as the minimum number compatible with a plurality. A premise of the form:

No A are B

has a model of the form:

[a] −b
[a] −b
[b]
[b]

In principle, the fleshing out of As as not Bs could occur later, but the model reflected our intuition that the fleshing out is immediately accessible to reasoners. The fact that a set is exhaustively represented does not prevent its co-occurrence with other tokens. The preceding model may be extended as follows, for example,

[a] −b
[a] −b
[b] c
[b] c

What cannot happen, however, is that Bs are represented as As, because the As in the model have been represented in their entirety. A premise of the form:

Some A are not B

has a model of the form:

a −b
a −b
b
b

where, again, the subject term, A, has been fleshed out explicitly as not B. The model supports the invalid converse conclusion, Some B are not A, because many people make this inference. The inference is invalid, and some critics have argued that the representation is, therefore, logically incorrect (see e.g., Ford, 1995). In fact, the invalid converse conclusion can be refuted by constructing a counterexample:

a −b
a −b
a b
a b

and so the theory is not committed to an irretrievable error. The representation of the premises is consistent and is based on simple psychological principles. It has only one major problem: it is wrong, as our present results will show.

The separate models of the two premises are combined by forming identities between the two sets of entities representing the middle term. Thus, given the premises

Some A are B.

All B are C.

the program constructs the model:

a [b] c
a
[b] c

Because there are always the same numbers of tokens of the middle term, this operation is simple, otherwise the program would have to ensure that there were equal numbers. The procedure produces a single, integrated model from any pair of syllogistic premises—granted, of course, that they have a middle term in common.

The program contains a procedure that formulates a conclusion; that is, it describes the relation between the end terms that holds in the models. Where there are no negative tokens in a model, the program formulates the conclusion holding between the two end terms, X and Y, in the following way: if each X in the model is a Y, then it concludes *All X are Y*; if at least one X in the model is a Y, then it concludes *Some X are Y*; otherwise, it responds that there is *No valid conclusion*. Where there are negative tokens in a model, if the Xs and Ys are disjoint and both are exhausted, or one of them and the middle term are exhausted, then the program concludes *No X are Y*. If at least one X is not Y, then it concludes *Some X are not Y*; otherwise, it responds that there is *No valid conclusion*. The program draws two conclusions from the initial combined model of the two premises,

scanning the model first in one direction and then the other. Given the preceding model, for example, the program draws the two conclusions:

Some A are C.

Some C are A.

Finally, the program tries to refute conclusions by constructing counterexamples, i.e., alternative models of the premises in which the conclusion is false. It is this component of the program that constructs multiple models of certain syllogisms. It uses three operations to search for models that refute conclusions, breaking entities into two, adding new entities to the model, or joining two separate entities into one. These operations, however, are constrained by the nature of the existing models of the premises. The first operation applies to models that contain, for instance, an entity of the form:

a b c

Because b is not exhausted, the entity can be broken into two separate entities without violating the meaning of the premises:

a b

b c

The second operation applies to end terms that are not exhaustively represented. It adds further instances of the end term to form a new model, e.g.,:

[a] [b] c becomes: [a] [b] c

[a] [b] c [a] [b] c

c

The third operation applies to end tokens that are not linked to the middle term or that can be made so by breaking an entity into two. It forms a new model by moving them to join up with the other end tokens, e.g.,:

[a] −b becomes: [a] −b [c]

[a] −b [a] −b [c]

[b] −c [b] −c

[b] −c [b] −c

[c]

[c]

To illustrate the program, we will use three examples that illustrate the main types of syllogisms. The first syllogism is based on the premises:

All A are B.

All B are C.

The program constructs the model:

[a] [b] c

[a] [b] c

It draws a conclusion interrelating A to C, All A are C. And it draws a conclusion interrelating C to A, All C are A. It next constructs an alternative model by adding a token of an unexhausted end term:

[a] [b] c
[a] [b] c
c

The conclusion, All A are C, still holds; it requires only one model, but the new model refutes the converse conclusion, and the program draws a new conclusion interrelating C to A, Some C are A. In fact, reasoners are governed by a figural effect and almost invariably draw a conclusion interrelating A to C.

The second example is based on the premises:

Some A are B.

No B are C.

The program constructs the initial model:

a [b] −c
a
[b] −c
[c]
[c]

from which it draws the conclusions:

No A are C.

No C are A.

It moves an end token, a, to create a second model of the premises:

a [c]
a [b] −c
[b] −c
[c]

This model refutes both the previous conclusions, and supports instead:

Some A are not C.

Some C are not A.

The program adds a new end token, a, to create a third model of the premises:

a [c]
a [b] −c
[b] −c
a [c]

The conclusion, Some A are not C, survives unscathed, but its converse is refuted, and so reasoners who formulated the converse will infer wrongly that there is no valid conclusion. No further models of the premises are possible, and so the correct valid conclusion is:

Some A are not C.

The third example is based on the premises:

All A are B.

Some B are not C.

The program constructs the initial model:

[a]	b	−c
[a]	b	−c
		c
		c

from which it draws the conclusions:

Some A are not C

Some C are not A.

It breaks the initial pair of entities into two and then moves the two tokens of c to create a new model of the premises:

[a]	b	c
[a]	b	c
	b	−c
	b	−c

This model refutes both conclusions, and so the program responds: No valid conclusion. Henceforth, we refer to these three sorts of syllogisms as *one-model* problems, *multiple-model* problems, and *no-valid-conclusion* problems, respectively.

The program obviates the criticisms raised by Hardman (1996), but it makes the same predictions as does the theory outlined by Johnson-Laird and Byrne (1991). The first prediction is that one-model syllogisms should be easier than multiple-model syllogisms and no-valid-conclusion problems: reasoners should be faster and make fewer errors with one-model syllogisms. The second prediction is that the erroneous conclusions that reasoners typically infer should correspond to those supported by the initial models of multiple-model and no-valid-conclusion problems. These conclusions match the mood of at least one premise, and so the model theory provides an alternative explanation for the atmosphere effect. This explanation rests on reasoning rather than a purely superficial matching of the verbal forms of premises and conclusions. But, as a referee reminded us (see also, Stenning and Yule, 1997), one-model problems have a conclusion that matches the mood of at least one of the premises, whereas multiple-model problems do not. Could it be that this simple principle accounts for the differences in difficulty? In other words, reasoners merely draw conclusions that match the mood of one of the premises, and so they will be right with one-model syllogisms and wrong with multiple-model syllogisms (for versions of this hypothesis, see Chater & Oaksford, 1999; Martin Levine, personal communication, 1994; Wetherick & Gilhooly, 1990). This putative explanation, however, fails to account for several phenomena that corroborate the model theory. First, if reasoners merely responded according to the mood of the premises, then they should be unaffected by whether a syllogism has one model or several. In fact, they are more likely to draw a conclusion matching the mood of a premise for one-model problems than for multiple-model problems (see Johnson-Laird & Byrne, 1991). Second, if reasoners are governed by the mood of the premises, they should never respond that nothing follows from the premises. In fact, they are more likely respond "nothing follows" for multiple-model problems than for one-model problems (see Johnson-Laird & Byrne, 1991). Third, according to the model theory, the quantifier "only" is implicitly negative, and so, inferences based on it should in general be harder, and reasoners should eschew conclu-

sions containing "only" in favor of those based on "all". Experiments have corroborated these predictions; in particular, reasoners seldom draw a conclusion containing "only," even when one or both premises contain it (Johnson-Laird & Byrne, 1991). We conclude that prima facie individuals reason from syllogistic premises, which they represent in the form of models. Readers may wonder, however, why the implementations of the model theory keep changing over the years. The answer will be revealed by the present investigations.

III. TWO STUDIES OF THE SPONTANEOUS USE OF DIAGRAMS

Experiments 1 and 2

Some theorists, as we have seen, argue that individuals not trained in logic rely on a mental equivalent of Euler circles (see e.g., Erickson, 1974; Ford, 1995; Stenning & Oberlander, 1995). In a recent, unpublished study of reasoning based on sentential connectives, Savary and Johnson-Laird discovered that none of the logically untrained participants could immediately externalize his or her mental representations, but that some of them developed skilled systems of diagrams isomorphic to mental models. Hence, the aim of our first two experiments was to examine what diagrams, if any, individuals not trained in logic would draw as they attempted to make syllogistic inferences. In Experiment 1, the participants were given paper and pencil, which they were free to use, and their task was to write down their conclusions to 20 pairs of syllogistic premises. In Experiment 2, the participants carried out the same task, but in addition, they were asked to "think aloud". The experiment anticipated Evans and Over's (1996) criticism that there have been too few studies of the model theory in which introspections were systematically recorded. But, our participants had great difficulty in thinking aloud as they used paper and pencil. Their protocols revealed little about the process of reasoning. We, therefore, report the two experiments together.

Method

Design and Materials. The participants carried out a set of 20 syllogistic inferences, which we selected to be representative of the total set of 64 possible pairs of premises. There were three sorts of problems:

4 one-model syllogisms.

8 multiple-model syllogisms.

8 no-valid-conclusion syllogisms.

Seven of the syllogisms were Figure 1 (see Introduction), six of Figure 3, and seven of Figure 4. We deliberately avoided using Figure 2 because the premises in this figure are logically equivalent to those in Figure 1, i.e., Figure 2 is obtained merely by swapping round the order of the two premises. Table 1 presents the 20 syllogisms, their mental models, the predicted responses based on their mental models, and the correct responses (in all capital letters). Both experiments were carried out in Italian, and the contents of

TABLE 1
The 20 Syllogisms in Experiments 1 and 2, Their Mental Models as Generated by the Computer Program, and Both the Predicted and Correct Responses (in CAPITAL Letters) Based on These Models

A. One-model syllogisms 1 All A are B All B are C ALL A ARE C SOME C ARE A	2 All B are A Some B are C SOME A ARE C SOME C ARE A
3 No A are B All C are B NO A ARE C NO C ARE A	4 Some B are A All B are C SOME A ARE C SOME C ARE A
B. Multiple-model syllogisms 5 Some A are B No B are C No A are C No C are A Nvc SOME A ARE NOT C	6 No A are B Some C are B No A are C No C are A Some A are not C Nvc SOME C ARE NOT A
7 No A are B All B are C No A are C No C are A Some A are not C Nvc SOME C ARE NOT A	8 All B are A All B are C All A are C All C are A SOME A ARE C SOME C ARE A

TABLE 1
Continued

9 No A are B [a] −b [a] −b c [a] −b c Some B are C [a] −b [a] −b [a] −b c [b] c [b] c [b] c [b] [b] [b] No A are C c c c No C are A Some A are not C Nvc SOME C ARE NOT A	10 All B are A a [b] −c a [b] −c a [b] −c No B are C a [b] −c a [b] −c a [b] −c [c] a [c] a [c] No A are C [c] [c] a [c] No C are A Some C are not A Nvc SOME A ARE NOT C
11 All A are B [a] b [a] b c Some C are not B [a] b [a] b c Some A are not C −b c −b c Nvc −b c −b c SOME C ARE NOT A	12 Some B are not A −a [b] c −a [b] c All B are C −a [b] c −a [b] c Some A are not C a a c Nvc a a c SOME C ARE NOT A
C. Syllogisms with no valid conclusions interrelating their end terms 13 All A are B [a] b −c [a] b c Some B are not C [a] b −c [a] b c c b −c Some A are not C c b −c Some C are not A NVC	14 All A are B [a] b c [a] b Some C are B [a] b b c c [a] b Some A are C c Some C are A NVC
15 Some A are B a b c a b Some B are C a b c b a Some A are C c b Some C are A c NVC	16 Some A are B a b [c] a b All C are B a b b [c] a Some A are C b [c] Some C are A NVC

TABLE 1
Continued

17 Some A are B a b −c a c Some B are not C a a b c b −c b −c c b −c c Some A are not C Some C are not A NVC	18 Some B are not A −a b c a c Some B are C −a b a b c a −a b a −a b c Some A are not C Some C are not A NVC
19 All A are B [a] b [c] [a] b [a] b All C are B [a] b [c] b [c] b [c] [a] b [c] [a] b All A are C b [c] All C are A Some A are C Some C are A NVC	20 Some B are not A −a [b] −c a [c] a [c] No B are C −a [b] −c −a [b] −c a [c] a −a [b] −c −a [b] −c No A are C a a −a [b] −c No C are A [c] [c] Some A are not C [c] Some C are not A NVC

Note: Nvc denotes the response "there is no valid conclusion." Problem 1 calls for two models to draw the conclusion, SOME C ARE A (see text).

each pair of premises were common hobbies, such as *singer* and *skier*, for the end terms, and common jobs such as *lawyer* and *baker*, for the middle terms.

Procedure. The participants were tested individually in a quiet room. They were told that they were taking part in an experiment on how people reason. Their task was to consider a series of problems, and, for each of them, to draw a conclusion that had to be true given that the premises were true. If they thought that there was no valid conclusion, then they were to write "no valid conclusion". They were also told that they were free to use the paper and pencil to help them, and that in any case the experimenter would make an anonymous video recording of what they wrote or drew. The participants in Experiment 2 were given an additional task, to try to think aloud as they tackled each problem. If they were silent for more than 5 s, the experimenter exhorted them to think aloud. We were interested in their spontaneous thoughts rather than their justifications for their conclusions, and so we did not ask them to explain or to justify their performance (cf. Ford, 1995). Each of the participants were given two three-term series problems as practice problems (one with a valid conclusion, and one without a valid conclusion). The practice problems were also used to convey to the participants that their conclusions should relate the two end terms. The video camera was in a fixed position so that it was in front of the participants, but above their normal point of view. It was focused on the area of the desk where the participants would write or draw on the paper. Each pair of premises was printed on a separate page of paper, and the participants wrote their answers beneath the premises. They were told that each term referred to a set of individuals in the situation. They could take as much time as they needed for each problem, but they were not allowed to return to an earlier problem in the sequence. After the participants had asked any questions about the task, and were certain that they understood it, they began the experiment proper. At the end of the experiment, the experimenter asked a series of questions about the participant's performance, including their knowledge of Euler circles.

Participants. We tested two separate sets of 20 volunteers in the two experiments. They were undergraduate students of psychology at the University of Turin, and none of them had received any formal training in logic. There is no selection procedure for admission to the University, and so the sample is closer to a sample of the general public of young adults than to university students in the English speaking world. Each experiment lasted for about half an hour.

Results and Discussion

Table 2 presents the 20 syllogisms that were used in both experiments and the percentages of the main conclusions that were drawn in the two experiments. As is evident, there were no major differences between the experiments in the percentages of correct responses. In what follows, we will analyze the results of the experiments in three main sections. First, we evaluate the predictions of the model theory. Second, we consider the systematic errors in reasoning. Third, we analyze what the participants' diagrams and their "think aloud" protocols revealed about their inferential strategies.

TABLE 2
The main conclusions and their frequencies in Experiment 1 (NoT, i.e., no ''think aloud'' procedure) and Experiment 2 (Tal, i.e., ''Think aloud'' procedure).

A. The main conclusions for the four one-model syllogisms

1	NoT	Tal	2	NoT	Tal
All A are B All B are C			All B are A Some B are C		
ALL A ARE C	18	13	SOME A ARE C	10	13
SOME A ARE C		3	SOME C ARE A	4	2
Nvc	1	3	Nvc	4	2
3	NoT	Tal	4	NoT	Tal
No A are B All C are B			Some B are A All B are C		
NO A ARE C	15	13	SOME A ARE C	6	8
NO C ARE A		2	SOME C ARE A	10	7
Nvc	4	4	Nvc	2	3

B. The main conclusions for the multiple-model syllogisms

5	NoT	Tal	6	NoT	Tal
Some A are B No B are C			No A are B Some C are B		
No A are C	6		No A are C	7	2
No C are A	1	2	No C are A	2	2
Nvc	5	7	Some A are not C	—	—
SOME A ARE NOT C	7	9	Nvc	11	9
Some A are C	1	2	SOME C ARE NOT A		5
7	NoT	Tal	8	NoT	Tal
No A are B All B are C			All B are A All B are C		
No A are C	11	14	All A are C	9	5
No C are A	4	3	All C are A	—	—
Some A are not C	1		SOME A ARE C	5	3
Nvc	3	2	SOME C ARE A	—	—
SOME C ARE NOT A	1		Nvc	6	10
9	NoT	Tal	10	NoT	Tal
No A are B Some B are C			All B are A No B are C		
			No A are C	9	6
No A are C	5	9	No C are A	2	3
No C are A	3	1	Some C are not A	1	
Some A are not C	—	—	Nvc	8	5
Nvc	9	7	SOME A ARE NOT C		3
SOME C ARE NOT A	2	2	Some A are C		2

TABLE 2
Continued

11	NoT	Tal	12	NoT	Tal
All A are B Some C are not B			Some B are not A All B are C		
Some A are not C	—	—	Some A are not C	2	
Nvc	9	12	Nvc	3	7
SOME C ARE NOT A	8	3	SOME C ARE NOT A	9	9
Some A are C		3	Some A are C	2	3
Some C are A	3		Some C are A	4	

C. The main conclusions for syllogisms with no valid conclusions interrelating their end terms

13	NoT	Tal	14	NoT	Tal
All A are B Some B are not C			All A are B Some C are B		
Some A are not C	13	15	Some A are C	2	2
NVC	4	1	Some C are A	8	8
Some A are C	2	1	NVC	8	7
15	**NoT**	**Tal**	**16**	**NoT**	**Tal**
Some A are B Some B are C			Some A are B All C are B		
			Some A are C	10	6
Some A are C	12	11	Some C are A		3
NVC	7	7	NVC	9	9
17	**NoT**	**Tal**	**18**	**NoT**	**Tal**
Some A are B Some B are not C			Some B are not A Some B are C		
			Some A are not C	2	
			Some C are not A	1	4
Some A are not C	8	9	NVC	14	13
NVC	8	5	No A are C		2
Some A are C	1	5	No C are A	2	1
19	**NoT**	**Tal**	**20**	**NoT**	**Tal**
All A are B All C are B			Some B are not A No B are C		
All A are C	5	3	No A are C	2	2
All C are A	—	—	No C are A	1	2
Some A are C	—	—	Some A are not C	2	2
Some C are A	—	—	Some C are not A	—	—
NVC	15	14	NVC	15	9

Note: The responses in CAPITAL letters are correct; the lowercase responses above them are predicted by the model theory; and responses in lowercase underneath them are not predicted by the model theory. Nvc denotes the response ''there is no valid conclusion.'' We include unpredicted responses only if there were made by at least two participants.

Experiment 1. The percentages of correct responses (and their mean latencies) for the three sorts of syllogism were as follows:

One-model problems:	80% correct	(26 secs)
Multiple-model problems:	21% correct	(41 secs)
No valid conclusion problems:	50% correct	(78 secs).

As the theory predicts, the one-model syllogisms yielded more correct conclusions than did the multiple-model syllogisms (Wilcoxon test $z = 3.92$; $p < .0001$). The theory also predicts that one-model problems should be easier than the problems that have no valid conclusion, although the nature of the response differs between the two sorts of problems. The comparison is not orthogonal to the previous one, but the difference was highly reliable (Wilcoxon test, $z = 3.36$; $p < .001$). The same patterns of reliability were also evident in the correct response times: one-model syllogisms yielded faster responses than did multiple-model problems (Wilcoxon test, $z = 2.48$; $p < .05$) or problems with no valid conclusions (Wilcoxon test, $z = 3.58$; $p < .0005$).

Experiment 2. The results of Experiment 2 showed the same patterns of performance as the first experiment. The percentages of correct responses (and their mean latencies) were as follows:

One-model problems:	76% correct	(31 secs)
Multiple-model problems:	21% correct	(51 secs)
No valid conclusion problems:	41% correct	(52 secs).

The one-model problems were reliably easier than both the multiple-model problems (Wilcoxon test, $z = 3.68$; $p < .0005$) and the problems with no valid conclusions (Wilcoxon test, $z = 2.72$; $p < .01$). The one-model problems were also solved faster than the multiple-model problems, although the difference was only marginally significant (Wilcoxon test, $z = 2.72$; $p < .06$), and were faster than the problems with no valid conclusions (Wilcoxon test, $z = 2.64$; $p < .01$). When the two experiments were grouped together, 39 out of the 40 participants performed more accurately with the one-model problems than with the multiple-model problems—a most robust result (Sign test, p was less than 1 in a million).

The participants had a mean of 8.5 correct answers; they made errors predicted by the model theory on a mean of 8.8 problems, and they made other errors not predicted by the model theory on a mean of 2.7 problems. Of the 40 participants, 36 made more predicted than unpredicted errors, three made more unpredicted errors than predicted errors, and there was one tie (Sign test, $p < .0001$). The model theory makes no predictions about errors on one-model problems. In every case, the errors made on these problems consisted of the response, no valid conclusion. We suppose that the participants either failed to construct an integrated representation of the premises or else guessed the response. Most of the unpredicted errors on multiple-model problems were also either "no valid conclusion" responses or Gricean implications from predicted conclusions. A typical example was that a syllogism with a valid conclusion of the form, Some of the A are not C, was taken to imply instead, Some of the A are C (see Grice, 1975). A theory can err by

predicting responses that participants do not make. As Table 2 shows, the model theory made few predictions of conclusions that participants failed to infer.

To determine the participants' strategies, we examined their protocols— any diagrams or words that they wrote in Experiments 1 and 2, and anything that they said in their "think aloud" protocols in Experiment 2. Of the 40 participants, 10 provided sufficient information for us to analyze. The remaining 30 participants either did not use the paper and pencil or did not say anything revealing in their "think aloud" protocols. In fact, the participants in Experiment 2 were singularly uninformative. Most merely read the premises aloud, sometimes repeated a premise, and then stated their conclusion. Here is a typical protocol for premises of the form:

All the B are A.

None of the B are C.

where, for simplicity, we use the terms A, B, and C, rather than the original names of occupations. The participant (Roberta) says:

> Some of these [pointing to A in the premises] are B. None of the B, which are A, are C. I don't know whether some A of the others, which are not B, can be C. I know that none of the B,which are A, are C. There's no conclusion.

The participant seems to have in mind the possibility represented in the following model:

```
a    [b]    −c
a    [b]    −c
            [c]
            [c]
```

and then to have realized that there can be As that are not Bs without being able to construct the model in which they are Cs. In any case, she was unable to formulate the conclusion, Some of the A are not C, which is consistent with such a model.

The strategies that we discerned in the data were mainly revealed by the participants' diagrams. But, it was difficult to categorize their performance, and we were unable to assign each participant to a single strategy. Indeed, the diagrams varied both between and within participants, and three participants appeared to change their strategies during the experiment. The simplest use of paper and pencil (2 participants) was merely to highlight the end terms in some way, such as drawing circles round them or drawing an arrow from the middle term to the end term in each premise. A slightly more complex procedure was used by one participant (Silvia, in Experiment 2) in which she added the following sort of annotation shown in bold, e.g.:

Some of the A are B = **C**

All C are B

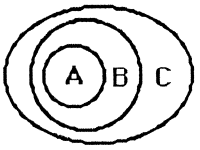
and wrote, "There is the same relation". She then drew the invalid conclusion: Some of the A are C.

Six out of the 40 participants used some form of Euler circles for at least one problem, although none of them adopted the circles for all the problems. They tended to construct just a single diagram, even for premises that could in principle be represented by several distinct diagrams. For example, two participants represented the premises:

All the A are B.

All the B are C.

with the following diagram:

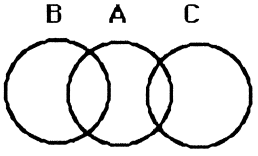


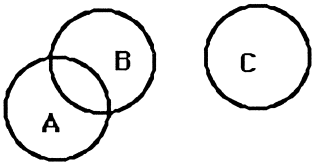
When more than one participant drew a diagram for the same syllogism, they sometimes drew different diagrams. Thus, the premises:

Some of the A are B.

None of the B are C.

elicited the following two distinct diagrams from two participants:



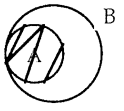


One participant performed as follows with the syllogism:

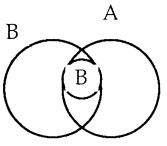
Some of the A are B.

None of the B are C.

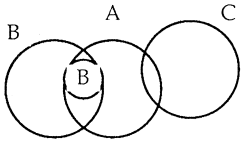
First, she drew a diagram of the first premise



Then for the same premise she drew a new diagram, where the two circles intersected, but their intersection contained another circle representing B:



Finally, she added a circle representing C:



In contrast, with the premises:

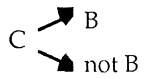
All the A are B

Some of the B are not C

this same participant drew an arrow to represent the first premise as follows:

A→B

Then, for the second premise, she drew a diagram of the following sort:



In general, the participants who used Euler circles drew one and the same diagram for a problem, although all of them sometimes used additional annotations, such as an equal sign between circles to indicate that they were equivalent. One participant even added dots within a circle to represent particular individuals.

Three participants devised their own symbolic systems for a few problems. For example, one participant represented the premises:

All the A are B.

All the B are C.

by using arrows

$A \rightarrow B \rightarrow C$

Another used equals signs:

$A = B = C$

And one represented the premises:

None of the A are B.

All the B are C.

with the combination:

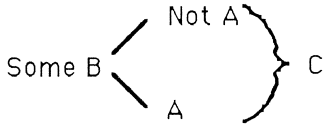
$A \neq B \rightarrow C$

and then drew the invalid conclusion, None of the A are C. Sometimes, these participants used a mixture of lines and labels as in the following examples. The premises:

Some of the B are A.

All the B are C.

were represented as:

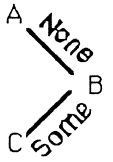


The premises:

None of the A are B.

Some of the C are B.

were represented as:



Finally, two participants used individual tokens to represent different sorts of individuals. One of these participants, for example, represented the premises:

Some of the B are not A.

Some of the B are C.

by enumerating the different sorts of possible individuals:

B not A
B A
B C

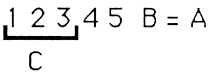
and correctly concluded that there was no valid conclusion. Another participant enumerated five possible instances of the middle term:

1 2 3 4 5 B

and represented the first premise, All the B are A, by adding the identity:

1 2 3 4 5 B = A

Next, this participant added an annotation to represent the second premise, Some of the B are C:



The participant then drew the correct conclusion:

Some of the A are C.

To the questions at the end of the experiment, all the participants responded that they had not heard of Euler circles. When the experimenter showed them examples of Euler circles, they also denied that they had encountered them before the experiment.

In summary, the experiment corroborated the predictions of the model theory about the relative difficulty of different syllogisms and about the main sorts of error that reasoners commit. Although the majority of participants provided no evidence about their strategies, those who did use paper and pencil relied on a variety of different diagrammatic techniques. Six of the participants used Euler circles for at least some of the problems, but we observed a variety of other notational devices.

IV. THE ROLE OF ALTERNATIVE MODELS IN SYLLOGISTIC REASONING

The evidence about syllogistic reasoning consists of the conclusions that reasoners draw—particularly those they draw for themselves, the latencies of their responses, and their introspections and use of diagrams. This evidence fails to pin down either the mental representations or the processes that underlie reasoning. Hence, one major unresolved question concerns a central principle of the model theory: the principle that reasoners search for alternative models of the premises. In some domains of reasoning, reasoners can construct all the possible models as they interpret each premise, e.g., for simple propositional, spatial, and temporal inferences. In reasoning that hinges on quantifiers, however, individuals are unlikely to be able to construct all the possible models as they proceed through the premises. Quantified assertions do not wear their logical hearts on their sleeves. For example, the assertion:

All of the actuaries are not businessmen

is ambiguous with respect to the scope of negation. It can be paraphrased either as:

None of the actuaries is a businessman

or as:

Not all the actuaries are businessmen.

This second interpretation is itself referentially indeterminate, as are most syllogistic premises. The model theory, therefore, postulates that reasoners begin by considering just a single model of the premises. In principle, they can then search for alternative models of the premises. Thus, the program formulates a conclusion based on the initial model and searches for a model that is a counterexample to the conclusion. If there are no alternative models, then a conclusion based on the initial model is valid. If there are alternatives, then any valid conclusion must be based on all the models. Hence, where the models have nothing in common, there is no valid conclusion interrelating the end terms in the premises (apart from weaker conclusions about possibilities).

Only conclusions that have no counterexamples are valid. Earlier studies suggested that people do not always put this principle into practice. Thus, some poor reasoners almost always draw a conclusion to any syllogistic premises (see Johnson-Laird, 1983, p. 120). They fail to search for alternative models of the premises and instead base their conclusions on their initial models. Other poor reasoners almost invariably respond that "nothing follows" from multiple-model problems. The natural explanation of their performance is that they construct alternative models, which they take as proof that there is no valid conclusion. They fail to discern that alternative models may have in common some relation between the end terms.

Polk and Newell (1995) proposed a radical alternative theory of model-based reasoning. They argue that reasoning is a largely verbal process, and they have implemented a computer program called VR (for Verbal Reasoning) that constructs mental models from syllogistic premises and either formulates a conclusion from them or declares that nothing follows. The interpretation of premises and the formulation of conclusions are indeed verbal processes that depend on both syntax and semantics. According to Polk and Newell (1995; see their Figure 5, p. 539), given the premises

Some B are A

All B are C

their VR program constructs the following initial model:

B′ C

B′ A C

where the apostrophe designates an "identifying" property, i.e., one that is more accessible because it derives from the topic or subject of a premise. This property is tried first in generating putative conclusions:

Some B are A

All B are C

These conclusions are not legal because they fail to interrelate the end terms. The program then repeatedly re-encodes the premises, first attempting to extract information about C from each of them and then attempting to extract information about A. Ultimately, the re-encoding of the first premise yields information about A, i.e., Some A are B, and so it can now construct an augmented model in which A is marked as an identifying property:

B′ C
A′
B′ A′ C

So, after generating the putative conclusions, Some B are A, All B are C, and Some A are B, from an earlier model, the program at last generates the legal conclusion:

Some A are C

from its final model. Only if the program fails to generate a legal conclusion from its re-encodings of the premises does it respond, No valid conclusion. Hence, Polk and Newell (1995) argue that the linguistic processes of encoding and re-encoding are central in deduction and that processes that are devoted exclusively to reasoning, such as searching for alternative models, play a smaller role. It is important to note, however, that the program constructs an alternative model in the example above, albeit to draw a legal conclusion, whereas our program constructs just a single model for this syllogism (see Table 1A). Polk and Newell (1995) do not rule out the possibility of the search for alternative models or of refutation. "The point is," they write, "that syllogism data can be accurately explained without positing a falsification strategy, in keeping with the hypothesis that such a reasoning-specific strategy is less important than verbal processes in explaining deduction." (p. 553).

If reasoning is a formal process based on rules of inference of the sort proposed by Braine and Rumain (1983), and Rips (1994), then counterexamples can play no role in reasoning because these rules make no reference to them. According to these theories, reasoners respond, Nothing follows, only if they fail to derive a conclusion from the premises. If reasoning depends on formal rules, then the believability of a conclusion cannot affect the process. But, if reasoning depends on models, then content can have direct effects on the process itself. In fact, when an initial conclusion is highly believable, reasoners tend not to search assiduously for a potential counterexample; but, when an initial conclusion is highly unbelievable, they tend to search harder for a counterexample (Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird & Garnham, 1989).

If reasoners search for counterexamples on occasion, then how does the search proceed? Martín-Cordero and Gonzàlez-Labra (1994) argued that a major flaw in the model theory is its failure to answer this question. The computer program that we described earlier contains a search procedure. The problem is, therefore, in the lack of evidence about how people search for counterexamples. The aim of Experiment 3 was to obtain such evidence.

Experiment 3: The Search for Counterexamples

Experiment 3 examined the competence of individuals not trained in logic to search for counterexamples to putative conclusions. If people are unable to refute conclusions in this way, then Polk and Newell (1995) are certainly correct in arguing that refutations play little or no role in syllogistic reasoning. The experiment was designed to externalize the process of searching for counterexamples. The participants were given complete syllogisms, and they had to try construct external models of the premises that refuted the

conclusions. In this way, the results should enable us to establish a theory of how individuals search for counterexamples.

Method

Design. The participants' task was to try to refute the conclusions of 20 syllogisms. They had to construct external models of the premises in which the given conclusions were false. These external models took the form of simple cut-out shapes representing three sorts of individuals, which the participants could use to depict the premises. The syllogisms were based on the same 20 pairs of premises used in the preceding experiments, but the premises were combined with conclusions. In the case of the four one-model premises, the conclusions were valid, and so it was impossible to refute them. In the case of the eight multiple-model problems and the eight problems with no valid conclusions, the conclusion was the one that is supported by the first integrated model generated by the computer program (for the premises and their mental models, see Table 1). The conclusions were, therefore, at least superficially plausible (as shown by the errors made by the participants in Experiments 1 and 2), but they could be refuted by constructing a model that is an example of the premises but a counterexample of the conclusion. Each participant in the experiment carried out all 20 problems in a different, random order.

Materials. The 20 syllogisms and their putative conclusions are summarized in Table 3. The premises were identical in form to the problems in Experiments 1 and 2, but the present experiment was carried out in English. The putative conclusions were supported by the initial models of the premises (see Table 1), and so they were valid for the one-model problems, but invalid for the multiple-model problems and the problems with no valid conclusions. Unlike the previous experiments, each syllogism concerned the same three occupations: cooks, musicians, and painters. This constraint was necessary to provide the participants with the materials for constructing external models. These materials consisted of simple cut-out paper shapes: a chef's hat to depict the chefs, a guitar to depict the musicians, and a palette to depict the painters. A simple stick figure depicted an individual, and the participants had six such figures, to which they could add hats, guitars, and palettes to represent a problem. Each problem was typed on a separate card, e.g.,

Premise 1: Some of the chefs are musicians.

Premise 2: All the painters are musicians.

Construct a picture that is not consistent with the conclusion: Some of the chefs are painters.

Procedure. The participants were tested individually in a quiet room. They were told that the experiment concerned reasoning about a series of problems, but that it was not a test of their intelligence. For each problem, they should carefully read the pair of premises and then try to construct a picture (from the cut-out shapes) consistent with both the premises, but not consistent with the stated conclusion. If they succeeded in this task, they had to respond: "This is a refutation". If they considered that it was impossible to depict

TABLE 3
The 20 syllogisms in Experiment 3

1 All A are B — one-model All B are C **(ALL A ARE C)** — 90%	2 All B are A — one-model Some B are C **(SOME A ARE C)** — 65%
3 No A are B — one-model All C are B **(NO A ARE C)** — 75%	4 Some B are A — one-model All B are C **(SOME A ARE C)** — 55%
5 Some A are B — multiple-model No B are C **No A are C** (SOME A ARE NOT C) — 80%	6 No A are B — multiple-model Some C are B **No A are C** (SOME C ARE NOT A) — 85%
7 No A are B — multiple-model All B are C **No A are C** (SOME C ARE NOT A) — 95%	8 All B are A — multiple-model All B are C **All A are C** (SOME A ARE C) — 40%
9 No A are B — multiple-model Some B are C **No A are C** (SOME C ARE NOT A) — 100%	10 All B are A — multiple-model No B are C **No A are C** (SOME A ARE NOT C) — 60%
11 All A are B — multiple-model Some C are not B **Some A are not C** (SOME C ARE NOT A) — 30%	12 Some B are not A — multiple-model All B are C **Some A are not C** (SOME C ARE NOT A) — 70%
13 All A are B — no-valid conclusion Some B are not C **Some A are not C** — 25%	14 All A are B — no-valid conclusion Some C are B **Some A are C** — 65%
15 Some A are B — no-valid conclusion Some B are C **Some A are C** — 70%	16 Some A are B — no-valid conclusion All C are B **Some A are C** — 70%
17 Some A are B — no-valid conclusion Some B are not C **Some A are not C** — 20%	18 Some B are not A — no-valid conclusion Some B are C **Some A are not C** — 45%
19 All A are B — no-valid conclusion All C are B **All A are C** — 85%	20 Some B are not A — no-valid conclusion No B are C **Some A are not C** — 20%

Note: Syllogisms are stated with the conclusion to be refuted in **bold**, the correct conclusion in capitals, and with the percentages of correct refutations in the experiment. The models for these syllogisms are shown in Table 1.

the premises in a way that refuted the conclusion, then they should say so, i.e., "It is impossible". They were told that that each term referred to a set of individuals in the situation, and that an individual could have more than one occupation, e.g., an individual might be a chef and a painter; or a chef, a musician, and a painter. The experimenter explained that a video recording would be made so that there would be an anonymous record of the sequences of pictures that the participant constructed for each problem. The camera was fixed in front of the participants, but above their normal point of view. It was focused on the area of the desk on which the participants constructed their external models.

Participants. Twenty Princeton University students, who had no training in logic, took part in the experiment. They were paid $5 per hour to carry out the experiment, which lasted for approximately 30 min.

Results

We transcribed each protocol for each problem from the video recordings and noted the sequence of external models that the participants constructed and their responses to the problems. There was an enormous diversity in the ways in which the participants sought counterexamples to conclusions. This diversity had two main sources. Individuals varied both one from another, and from one trial to another, in how they interpreted the different sorts of quantified premise. As we have already mentioned, these assertions do not wear their logic on their sleeves, and the participants varied in the models that they constructed from them. Individuals also varied, again both one from another and from one trial to another, in their overall strategy of searching for counterexamples. Our analysis of the results will accordingly be in three parts. First, we deal with the relative difficulty of the problems and the causes of error; second, we describe the participants' initial models of the premises; and, third, we assess the variation in their strategies.

We counted as correct counterexamples those cases in which a participant constructed an external model that satisfied the premises, but refuted the conclusion, and declared that it was a refutation. As a correct response to the one-model problems, we counted those cases where a participant constructed an external model that satisfied the premises and declared that the task of refutation was impossible. Table 3 summarizes the percentages of correct refutations for the 20 syllogisms. The overall percentage of correct responses in cases where a conclusion could be refuted was 59%; and the overall percentage of correct responses in cases where a conclusion could not be refuted was 71%. Each participant was able to refute putative conclusions, and the range in performance was from 95% correct responses by the best participant to 25% correct responses by the poorest participant. It is difficult to assess the probability of making a correct refutation by chance. It requires, where possible, the construction of a model of the premises in which the conclusion is false, and then the statement of the correct response. If we assume, conservatively, a chance probability of 1 in 10, i.e., a probability of 1/5 of constructing the correct model and a probability of 1/2 of making the correct response, then 19 participants performed better than chance, and there was one tie ($p = .5^{19}$, i.e., less than 1 in half a million).

There were differences in the difficulty between the three sorts of problems. The one-model problems (71% correct) were reliably easier than the other two sorts of problem [15 participants performed better with one-model problems, three did not, and there was one tie (Sign Test: $p < .004$)]. However, the responses to one-model problems ("impossible to refute") differ in kind from the responses to the other two sorts of problem ("this is a refutation"). Although there was no reason to predict the difference, the multiple-model problems (70% correct) were reliably easier to refute than were the problems with no valid conclusions (50% correct). Sixteen participants performed better with multiple-model problems, two did not, and there was one tie (Sign Test: $p < .001$).

A major cause of error, and thus of the differences in difficulty among the three sorts of problem, was the mood (see Introduction) of the putative conclusion. Overall, the participants were correct on the following percentages of problems:

A conclusions: 72%
I conclusions: 66%
E conclusions: 82%
O conclusions: 35%

and the effect of mood was reliable (Friedman two-way non-parametric analysis of variance, Fr = 27,045; p < .0001). Moreover, all the participants performed worst with O conclusions (Sign Test: $p = .5^{20}$). In general, they grasped that to refute a conclusion in the A mood (all the A are C), they needed to construct a model in which not all of the A are C; and they grasped that to refute a conclusion in the E mood (none of the A are C), they needed to construct a model in which some of the A are C. With conclusions in the O mood (Some of the A are not C), however, they often constructed a model in which some of the A were C, and they sometimes constructed a model in which none of the A were C. The correct counterexample calls for a model in which all the A are C. Likewise, the participants occasionally thought that they had refuted a conclusion in the I mood (Some of the A are C), when they had merely constructed a model in which some of the A were not C.

Another cause of errors was the interpretation of premises in the A mood. If such a premise is interpreted so that the set A is co-extensive with the set B, then errors are inevitable in some cases, but not in others. The co-extensive interpretation has no effect on one-model problems, which remain irrefutable. But, the co-extensive interpretation is likely to mislead reasoners when the subject of the first premise is the middle term, i.e., All the B are A, because then regardless of the mood of the second premise, the set of tokens of B in the model is unlikely to be altered. Consider, for example, the following problem:

All the B are A.
All the B are C.
∴ All the A are C.

If the participants built a model based on the co-extensive interpretation of the first premise:

b	a	c	or:	b	a	c
b	a	c		b	a	c
						c

TABLE 4
The percentages of the different sorts of initial models constructed by the participants in Experiment 3.

Moods of premise	A = B	A < B	B < A	A over B	A disj B
A: All A are B (8 sylls)	88	8	*4	—	—
I: Some A are B (5 sylls)	18	15	57	10	—
E: No A are B (4 sylls)	—	—	—	—	100
O: Some A are not B (3 sylls)	*7	—	67	5	21

Note: These data are based solely on those trials in which the participants constructed their initial model form the first premise in the statement of the problem. A = B, two sets are co-extensive; A < B, A is properly included within B; B < A, B is properly included within A; A over B, the two sets overlap one another; and A disj B, the two sets are disjoint with no members in common. * denotes erroneous model. A denotes either the end term or the middle term.

then it was difficult for them to falsify the putative conclusion. In contrast, consider the following problem:

All the A are B.
All the C are B.
∴ All the A are C.

With this problem, the Bs are not exhaustively represented, and so the participants should be more likely to realize that there could be Bs that are not As, which in turn could be co-extensive with the Cs. Indeed, the majority of the participants constructed this sort of external model. Only a few erred by making co-extensive interpretations of both premises.

The participants varied in how they interpreted the four different moods of the premises. Table 4 presents the percentages of the different kinds of initial models they constructed for each of the four moods. To minimize residual effects, these data come solely from those trials in which the participants constructed their initial model from the first premise in the problem. The second premise almost certainly influenced the interpretation of the first premise in those moods that are referentially indeterminate, i.e., the A, I, and O moods. The preferred interpretation for first premises in the A mood was the co-extensive one, e.g.:

a b
a b

for problems in the AA, AI, and AE moods. But, when the second premise was in the O mood, the participants were more inclined to build models in which the subject tokens were properly included within the object tokens, e.g.:

a b
a b
b

Evidently, a second premise, such as Some of the B are not C, focused reasoners' attention on Bs that were not As in their interpretation of the first premise. An analogous pattern of influence occurred with first premises in the O mood. The prevalent interpretation for problems in the OI and OE moods was one in which the tokens of the predicate were treated as a subset of tokens of the subject, e.g., the premise, Some of the A are not B, was interpreted as:

a b
a b
a

But, when the second premise was in the A mood, the participants were more inclined to build models in which the two sets of tokens were disjoint:

a
a
b
b

The preferred interpretation for first premises in the I mood was also for the predicate tokens to be included within the subject tokens, e.g., Some of the A are B, was interpreted as:

a b
a b
a

for the IA, II and IO problems. But, the participants were more likely to make a co-extensive interpretation when the second premise was in the E mood.

The most striking aspect of the results was the great variety of the participants' strategies. They sometimes began by constructing a model of the first premise to which they added the information from the second premise; they sometimes proceeded in the opposite order. Sometimes, their initial model satisfied the conclusion, and so they modified the model to refute the conclusion. Sometimes, they constructed an initial model of the premises that immediately refuted the conclusion. Here, to convey the diversity of strategies, we will summarize performance with a typical problem (5):

Some of the A are B.
None of the B are C.
∴ None of the A are C.

Of the 20 participants, 16 correctly refuted the conclusion. Five of these participants constructed an initial model of the premises that was consistent with the conclusion:

a
a b
c

where we ignore throughout the actual numbers of tokens of each type that the participants constructed. Two of the five participants then refuted the conclusion by adding a C to an A (an operation that the computer program also carries out):

1. a c
   a b
   c

Another two of the five participants added an A to a C (which the program can also do):

2. b a b a
   b becomes: b
   c a c

The remaining participant of the five introduced a new B and an A, and added a C to the A:

3.
```
a   b                    a   b
a                        a   b
        c   becomes:         b
                         a       c
                                 c
```

In contrast, 11 of the 16 participants who were correct refuted the conclusion in their initial model of the premises. Six of them did so with the following model:

4.
```
a       c
a   b
```

And three of them did so with the model:

5.
```
a   b
a       c
    b
```

The other two out of the 11 built slight variants of the first of these models. Thus, the 16 participants reached a correct counterexample by using at least five distinct strategies.

We now compare the participants' strategies with those we implemented a priori in the computer program. The program uses three operations to search for counterexamples: 1) breaking individuals into two separate individuals, 2) adding a new token to an existing individual in the model, and 3) joining two separate individuals into one.

The strategy of breaking one individual into two was occasionally used by some participants. For example, with the following problem with no valid conclusion

Some of the A are B.
Some of the B are C.
∴ Some of the A are C.

Two participants constructed an initial model that satisfied the conclusion:

```
a
a   b
a   b   c
    b   c
```

and then broke the key individual (a, b, c) into two separate individuals (a, b) and (b, c) to refute the conclusion. One reason why this strategy might appear to be used rarely is that the participants may sometimes be carrying it out in their minds as they construct their initial models of syllogisms. Thus, in the case of the following problem with no valid conclusion:

All the A are B.
All the C are B.

Fourteen participants constructed the initial model:

```
a   b
a   b
    b   c
    b   c
```

which immediately falsifies the conclusion. Because most other problems elicited initial models consistent with the putative conclusion, the participants may have momentarily envisaged the following possibility:

```
a    b
a    b    c
     b    c
```

only to refute it as they constructed their external counterexample.

The most frequent strategy was to add a new individual to a model. It occurred, for example, in the refutation of the following problem:

All the B are A.
All the B are C.
∴ All the A are C.

Six participants correctly refuted the conclusion by adding an additional A to their initial model of the premises:

```
b    a    c                    b    a    c
b    a    c    becomes:        b    a    c
                                     a
```

The program uses the same method for this problem, but then adds an additional C to refute the conclusion, All the C are A:

```
b    a    c
b    a    c
     a
          c
```

It was not necessary to refute the latter conclusion in the experiment, yet one of the six participants took the same step. Two separate participants used the same general method, but began their model with the second premise:

```
b    c    a                    b    c    a
b    c    a    becomes:        b    c    a
                                          a
```

The strategy of joining two individuals to make a new individual also occurred in the experiment. For instance, with the following problem that has no valid conclusion

Some of the A are B.
Some of the B are not C.
∴ Some of the A are not C.

most participants erred because they failed to grasp that the refutation calls for a model in which all the A are C. Of the four participants who correctly falsified the conclusion, one constructed the following model that satisfies the conclusion, but not the premises:

```
a
a    b    c
     b    c
```

and then moved the C to the A:

a c
a b c
b

to make a correct counterexample.

There are major discrepancies between the computer program and the participants' performance. The most salient difference is that the program uses a single deterministic strategy. It tackles each problem in a fixed routine. The participants, however, appeared to be more flexible in their approach, and different individuals used different strategies for the same problems. They also use different sorts of model for the same individual premises. Still another difference concerns negation. Consider, for example, the following multiple-model problem:

None of the A are B.
All the B are C.
∴ None of the A are C.

Sixteen of the participants began with an initial model of the premises that satisfied the conclusion. They then refuted the conclusion either by adding a C to an existing A (10 participants):

a a
a becomes: a c
b c b c

or by adding Cs to all the existing As (3 participants):

a a c
a becomes: a c
b c b c

The program makes these two steps one after another. But, it makes explicit that the model of the premise, None of the A are B, has negative tokens:

[a] −b
[a] −b
[b]
[b]

The participants did not use any overt representation of the fact that As cannot be Bs, albeit, we did not provide them with any external way in which to represent negation. In our view, however, they are likely to use the printed premises as a reminder, or to make a "mental footnote" to the same effect. Indeed, given the one-model problem, in which the conclusion is impossible to refute:

None of the A are B
All the C are B
∴ None of the A are C

they tended to construct the following model:

a
a
c
c

Our first thought was that for some reason this problem elicited a model of the conclusion alone. At the end of the experimental session, the experimenter asked some of the participants why they had not represented the Bs. They explained that they had built a "picture" of the first premise, but that they did not need to represent the Bs, and that they then added the Cs referred to in the second premise.

Discussion

The participants were able to refute conclusions by constructing external models that were examples of the premises, but counterexamples of the conclusion. One source of difficulty, however, was that many participants had an improper grasp of what counts as a counterexample to a conclusion of the form:

Some of the A are not C.

They assumed that it sufficed to establish the alternative conclusion:

Some of the A *are* C.

Four of the eight problems with no valid conclusion were presented with a conclusion in the O mood, and this factor may explain why these problems were more difficult than multiple-model problems. The converse error also occurred, although to a lesser degree. In contrast, most participants grasped that a conclusion of the form:

All the A are C

was refuted by a model in which not all of the A are C (or, equivalently, some of the A are not C). Similarly, they grasped that a conclusion of the form:

None of the A are C

was refuted by a model in which some of the A are C.

According to the model theory, it is harder to envisage the circumstances in which an assertion would be false than to envisage the circumstances in which it would be true. Patricia Barres (unpublished data) corroborated this claim for assertions formed with different sentential connectives. But, why do -reasoners not trained in logic go wrong most often in falsifying assertions of the form, Some A are not C? One possibility is that they assume that a negative assertion is refuted if the corresponding un-negated assertion is true, i.e., they merely drop the word *not* from the assertion and seek to establish the truth of the resulting assertion. This strategy, of course, is not so readily applied to:

None of the A are C

because the result is not a grammatical sentence. A simple test of this idea would be to present participants with a putative conclusion of the form:

Not all of the A are C.

It is equivalent to the assertion, Some of the A are not C, but now the strategy of dropping the negative word *not* yields the correct refutation:

All of the A are C.

We have been careful hitherto to disclaim any psychological reality for the operations in the computer program. But, the operations that construct counterexamples turn out to be remarkably similar to the operations used by our participants. The crucial resemblance is that the participants tend to construct initial models that satisfy both the premises and

the conclusion. Once this step has been taken, the best way to refute an invalid conclusion is to modify the model, and there are only a few possible operations that will do so, e.g., breaking an existing individual into two, creating a new individual by adding a new token to an existing individual, and uniting two separate individuals into one. These three operations embodied in the program do occur. Other potential operations, such as the creation of a new individual by removing a token from an existing individual, did not occur in our experiment.

There are three major discrepancies between the program and the participants' performance. First, the program follows a deterministic strategy. Given a particular pair of premises, it always proceeds in the same way. Our participants, however, varied considerably one from the other in what they did, and they seemed likely to vary if they were to encounter the same problem twice (for evidence on this point, see Experiment 4 below, and Johnson-Laird and Steedman, 1978). Second, the program uses a fixed interpretation of the premises, whereas given a premise of a particular form, our participants sometimes created one sort of model and sometimes constructed another sort of model—a phenomenon that is much more in line with Polk and Newell's (1995) theory. Third, the program departs from human performance in its explicit representation of negation. Our participants, perhaps because they lacked any external symbols for negation, appeared to represent them only as "mental footnotes".

One moral of our results is that individuals not trained in logic are able to construct multiple models of syllogistic premises to refute conclusions. This ability is beyond the explanatory scope of all current formal rule theories, which refute conclusions merely by failing to find formal derivations of them. A critical issue, however, is whether individuals spontaneously use the same strategy when they have to draw syllogistic conclusions for themselves. To examine this issue, we carried out one final experiment.

Experiment 4: Reasoning with External Models

This experiment was designed to observe the external models that the participants built in drawing their own conclusions from syllogistic premises. For purposes of comparison, each participant also carried out the inferential task without being allowed to construct external models.

Method

Design. The participants acted as their own controls and drew their own conclusions from the same set of 48 syllogisms in two conditions, 1 week apart. Half the participants carried out the task first using external models and then without using them; and half the participants carried out the two conditions in the opposite order. The two sets of syllogistic premises had different contents, and the assignment of contents was also counterbalanced over the participants, so that in effect there were eight different groups of participants. The order of the problems in both conditions was random for each participant.

Materials. The problems were based on 48 pairs of syllogistic premises—all the possible premises in Figures 1, 3, and 4. We did not use Figure 2 because premises in this Figure are logically equivalent to those in Figure 1, differing only in the order of the two premises in each pair. The problems are shown in Table 5. They consist of seven one-model problems; 14 multiple-model problems, with valid conclusions; and 27 multiple-model problems, with no valid conclusions interrelating their end terms. The experiment was carried out in Italian with native speakers of that language. One set of contents concerned three sorts of individuals cooks (cuochi), musicians (musicisti), and painters (pittori); and the other set of contents concerned swimmers (nuotatori), photographers (fotografi), and farmers (contadini). These terms were chosen because they could be easily represented by pictures in the external model condition (cook's hat, guitar, and palette, respectively, for the first set; and underwater mask, camera, and basket of vegetables, respectively, for the second set).

Procedure. The participants were tested individually in a quiet room. After a preamble in which they were told that they were taking part in an experiment on how people reason, they were read the following instructions:

> Read carefully the premises that I'll present to you, and draw a conclusion from them. The assertions always concern three sorts of individuals: <term A - hobby>, <term B - job>, and <term C - hobby>. The conclusion must relate the terms not directly related in the premises, namely <term A - hobby> and <term C - hobby>.

Next, the participants were given two three-term series problems, one with a valid conclusion relating the end terms, the other with no valid conclusion, and they were invited to consider what, if anything, followed. The instructions continued:

> The problems you will deal with are slightly different from these. Indeed, for each problem, your response should be one of the following [the experimenter presented the five sorts of responses typewritten on a single sheet of paper]:
>
> All. . . are. . .
>
> Some. . . are. . .
>
> None. . . are. . .
>
> Some. . . are not. . .
>
> Nothing follows.
>
> Please note that a particular individual may be of more than one sort, e.g., he may be both a <term A - hobby> and a <term B - job>. Any conclusion that you draw must be true given that the premises are true.

In the external model condition, the experimenter gave the further instructions:

> Your task, in solving each problem, is to help yourself by using some shapes to construct a picture of the premises. In the picture you construct, you must represent the three sorts of individuals, <term A - hobby>, <term B - job >, and <term C - hobby>. Note that a particular individual may be of more than one sort. The picture

TABLE 5
The main conclusions and their frequencies in Experiment 4.

A. Conclusions for syllogisms of Figure 1

AB-BC	A			I			First premise E			O		
	□	NoE	Ext	□	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext
	All C are A		1				No A are C	16	11			
A	SOME C ARE A	1		SOME A ARE C	16	17	No C are A	1		Some A are not C	12	3
	ALL A ARE C	17	16	SOME C ARE A	2	1	Some A are not C		3	Some C are not A		2
	SOME A ARE C	2	1	All C are A		1	Nvc	3	3	NVC		1
	Some A are not C		1	Some A are not C	1		SOME C ARE NOT A		2	Some A are C	6	14
	Some C are not A		1	Nvc	1	1	Some A are C		1	Some C are A	2	
	◪◪	NoE	Ext	◪◪	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext
	Some A are C	14	8									
I	Some C are A	3										
	NVC	3	7				No A are C	3	6			
	Some A are not C		2				Some A are not C	2		Some A are not C	1	6
	Some C not A		1	Some A are C	7	6	Nvc	8	11	Some C are not A		1
	Some B not A		1	NVC	13	12	SOME C ARE NOT A	7	1	NVC	14	9
	No A are C		1	Some A are not C		2	Some A are C		2	Some A are C	5	4
	□	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
E							No A are C	6	3			
	NO A ARE C	18	18	No A are C	1	4	NVC	13	14	Some A are not C	7	5
	NO C ARE A		2	Nvc	6	5	All A are C	1	1	Some C are not A		1
	Some A are not C	1		SOME A ARE NOT C	11	9	Some A are not C		1	NVC	10	12
	Nvc	1		Some A are C	2	2	Some C are A		1	Some A are C	3	2
	◪◪	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
O							No A are C	2	4			
	Some A are not C	11	7	Some A are not C	3	3	No C are A	1				
	NVC	6	6	NVC	13	10	Some A are not C	2	2			
	Some A are C	2	5	Some A are C	3	6	Some C are not A	1		Some A are not C	6	3
	Some C are A		2	Some C are A	1		NVC	13	12	NVC	13	13
	All C are A	1		No A are C		1	Some A are C	1	1	Some A are C	1	4

TABLE 5
Continued

B. Conclusions for syllogisms of Figure 3

AB-CB	First premise: A			First premise: I			First premise: E			First premise: O		
A	◪◪	NoE	Ext	◪◪	NoE	Ext	□	NoE	Ext	□□	NoE	Ext
										Some C are not A		1
				Some A are C	8	9				Nvc	2	6
	All A are C	12	11	Some C are A	2	3				SOME A ARE NOT C	7	3
	Some A are C	1	4	NVC	9	3	NO A ARE C	17	18	Some A are C	8	8
	NVC	7	4	Some A are not C	1	4	NO C ARE A	3	1	Some C are A	3	1
	Some A are not C		1	No A are C		1	Nvc		1	No A are C		1
I	◪◪	NoE	Ext	◪◪	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext
							No A are C	2	3	Some A are not C		1
				Some A are C	1	4	Some A are not C		2	Some C are not A	1	
	Some A are C	3	8	Some C are A	1		Nvc	8	7	NVC	12	12
	Some C are A	9	5	NVC	17	14	SOME C ARE NOT A	9	6	No C are A		1
	NVC	8	6	Some A are not C	1	1	Some A are C		2	Some A are C	7	5
	Some C are not A		1	Some A are B		1	Some C are A	1		Some C are A		1
E	□	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
				No A are C	2	1						
				No C are A	1							
	NO A ARE C	12	10	Some C are not A	1	1				No A are C		1
	NO C ARE A	7	6	Nvc	10	10	No A are C	2	2	Some A are not C	3	1
	Some A are not C	1		SOME A ARE NOT C	5	6	NVC	18	15	NVC	13	12
	Some C are not A		1	Some A are C		1	Some A are C		2	Some A are C	4	5
	Nvc		3	Some C are A	1	1	Some A are not C		1	Some C are A		1
O	□□	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
	Some A are not C	2	4				No A are C		3			
	Nvc	9	6				All C are not A	1		Some A are not C	1	3
	SOME C ARE NOT A	4	1	Some A are not C		1	Some C are not A	2	1	NVC	14	13
	No A are C		5	NVC	18	15	NVC	12	12	Some A are C	5	2
	Some A are C	1	2	Some A are C	2	3	Some A are C	3	3	Some C are A		1
	Some C are A	4	2	Some C are A		1	Some C are A	2	1	No C are A		1

TABLE 5
Continued

C. Conclusions for syllogisms of Figure 4

BA-BC	First premise: A			First premise: I			First premise: E			First premise: O		
A	□□	NoE	Ext	□	NoE	Ext	□□	NoE	Ext	□□	NoE	Ext
										Some A are not C	3	1
										Nvc		3
	All A are C	12	7	SOME A ARE C	11	12				SOME C ARE NOT A	6	4
	All C are A		1	SOME C ARE A	7	5	No A are C	8	12	No A are C		1
	SOME A ARE C	6	5	All A are C		2	No C are A	4	2	All A are C	1	1
	SOME C ARE A		2	Some A are not C		1	Some A are not C	1		Some A are C	8	7
	Nvc	2	3	Some C are not A	1		Nvc	6	5	Some C are A	2	2
	Some A are not C		2	Nvc	1		SOME C ARE NOT A	1	1	Some B are C		1
I	□	NoE	Ext	◪◪	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext
							No A are C	3	6			
	SOME A ARE C	15	16	Some A are C	3	2	No C are A	1		Some A are not C	3	
	SOME C ARE A	2	2	NVC	15	17	Some A are not C		2	NVC	15	15
	Some A are not C	1		No A are C	1		Nvc	13	8	Some A are C	2	4
	Nvc	2	2	Some A are not C	1	1	SOME C ARE NOT A	3	4	Some C are A		1
E	□□	NoE	Ext	□□	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
				No A are C	2	5				No A are C	4	4
	No A are C	12	11	Nvc	13	9	No A are C	4	1	No C are A		1
	No C are A	1		SOME A ARE NOT C	5	4	NVC	15	18	Some A are not C	5	6
	Nvc	4	8	Some A are C		1	Some A are C		1	NVC	10	9
	SOME A ARE NOT C	3	1	Some C are A		1	Some A are not C	1		Some A are C	1	
O	□□	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext	◪◪	NoE	Ext
							No A are C	3	4			
	Some C are not A	1	1				No C are A	1	1			
	Nvc	2	3	Some A are not C	2	2	All C are not A	1		Some A are not C	—	—
	SOME A ARE NOT C	13	9	Some C are not A	1		Some A are not C	1		Some C are not A	—	—
	Some A are C	3	5	NVC	14	15	Some C are not A		2	NVC	14	17
	Some C are A		1	Some A are C	2	2	NVC	14	12	All A are C	1	
	All C are A	1	1	Some C are A	1	1	Some A are C		1	Some A are C	5	3

NoE, no external aids session; and Ext the external aids session. The responses in CAPITAL letters are correct; the lowercase responses above them are predicted by the model theory; and responses in lowercase underneath them are not predicted by the model theory. Nvc, the response ''there is no valid conclusion''. □, one-model problem; □□, multiple-model problem; and ◪◪, problem with no valid conclusion interrelating the end terms.

TABLE 6
Percentages of correct responses given to the syllogisms of Experiment 4, according to the number of models and experimetnal condition.

	No External Model	External Model
One model	93	89
Multiple model	29	21
No valid conclusion	60	57

in which you represent the premises should help you to determine what conclusion, if any, follows from the premises.

Participants

Twenty student volunteers from the University of Turin took part in the experiment. They had no training in logic, and they had not participated in any previous experiment on reasoning. Each session of the experiment lasted for approximately 1 hr.

Results

Table 5 presents the main conclusions and their frequencies for each of the 48 pairs of premises. There was no reliable effect on accuracy whether the participants constructed external models (51% correct overall) or did not construct external models (55% correct overall; Wilcoxon test, $z = 1.28$; $p > .1$). Likewise, there was no reliable difference between the first session of 48 syllogisms (51% correct) and the second session of 48 syllogisms (55% correct; Wilcoxon test, $z = 1.15$; $p > .1$). And there was no interaction between condition and session (Mann-Whitney test, $U = 26.5$; $p > .2$). However, there was one reliable difference between the two conditions. The participants drew a slightly more diverse set of conclusions (a mean of 4.3 different conclusions to each problem) when they constructed external models than when they did not (a mean of 3.6 different conclusions; Wilcoxon test, $z = 2.93$; $p < .01$). The participants were moderately consistent in the conclusions that they drew in the two conditions: 60% of their conclusions were logically identical, which is well above chance given that there are 9 possible responses to each syllogism. Table 6 presents the overall percentages of correct responses to the one-model, multiple-model, and no-valid-conclusion problems in the two conditions. As in the previous experiments, one-model syllogisms were reliably easier than multiple-model syllogisms both with external models (Wilcoxon test, $z = 3.92$; $p < .0001$) and without (Wilcoxon test, $z = 3.82$; $p < .0002$).

We scrutinized the external models that the participants constructed to determine whether they constructed a sequence of alternative models compatible with a search for counterexamples. We examined first the problems for which the participants constructed two or more distinct models, i.e., models containing different sorts of individuals, ignoring mere differences in the numbers of tokens. The participants constructed such multiple models on 39% of trials, and all 20 participants built them, ranging from two participants who built such sequences on 75% of problems, down to one participant who built them

on only 8% of the problems. The results corroborated a crucial prediction of the model theory: the participants were more likely to construct two or more models for the multiple-model problems (37% of such problems) and the problems with no valid conclusions (48% of such problems) than for one-model problems (11%). All 20 participants were in accord with this prediction (Sign test, $p = .5^{20}$). Overall, there were 18% of trials on which the participants constructed multiple models of the premises and then responded correctly.

The data in the previous paragraph almost certainly underestimate the construction of multiple models. When the participants constructed just a single model of multiple-model problems or problems with no valid conclusion, they often made *correct* responses that were not consistent with that model. For example, given the problem:

None of the A are B.

Some of the B are C.

a participant constructed the following model:

```
a
a
      b     c
      b
```

which supports the conclusion:

None of the A are C.

Indeed, several participants drew this conclusion from such a model. Yet, one participant instead drew the correct conclusion:

Some of the C are not A.

Such a case suggests that the participant imagined an alternative model:

```
a           c
a           c
      b     c
      b
```

which refuted the first conclusion. Similarly, given the problem:

All the A are B.

All the C are B.

a participant constructed the following model:

```
a     b     c
a     b     c
      b
```

which supports the conclusion

All the A are C.

The participant, however, correctly responded that there was no valid conclusion. Such cases suggest that reasoners are imagining an alternative model that refutes the conclusion:

```
a     b
a     b
      b     c
      b     c
```

These sorts of correct refutations occurred on 14% of trials.

Some participants drew a conclusion, and then without adding any further tokens to the model they changed their conclusion in a way that suggested that they were taking into account another model. For instance, given the premises:

Some of the A are B

None of the C are B

one participant built the model:

```
a     b
a
            c
            c
```

and concluded:

None of the A are C.

But then he changed his mind, and drew the conclusion:

Some of the A are not C.

Another tendency consistent with a search for counterexamples was apparent in the results of five of the 20 participants, who all on one or more occasions drew a conclusion based on their initial models, then they said "No", or "Wait", and then went on to add a new token to their models, and to revise their conclusions. For instance, given the premises:

None of the B are A

All the B are C

one participant constructed the model:

```
b     c
b     c
      c     a
            a
```

and drew the conclusion:

Some of the A are C.

He then removed a token of C to yield:

```
b     c
b     c
            a
            a
```

and drew a new conclusion:

Some A are not C.

He then remarked, "It's only possible. None of the A are C is only possible". Finally, he responded erroneously, "No valid conclusion."

One unexpected but systematic result occurred with the problems that have no valid conclusion. We can divide these problems into those based on two affirmative premises and those based on at least one negative premise. If the participants reached the correct conclusion (nothing follows) for an affirmative problem, they tended to do so by con-

structing a single model that refuted any affirmative conclusion (56% of occasions). For example, given:

Some of the A are B.

Some of the B are C.

a participant constructed the following model:

```
a    b
a
     b    c
     b
```

and responded, "Nothing follows." In contrast, if the participants reached the correct conclusion for a problem with at least one negative premise, then they tended to do so either by denying the obvious conclusion supported by the single model that they had constructed (59% of occasions) or by constructing at least two alternative models that refuted the conclusion (35% of occasions). Eighteen participants refuted both affirmative and negative problems, correctly responding, "Nothing follows", and all of them fit this pattern (Sign test, $p = .5^{18}$). Our interpretation of this difference is that reasoners were able to construct an immediate counterexample to any conclusion suggested by the affirmative problems, but they had at least to construct one model of the negative problems before they could envisage a counterexample, either an external model or one in their mind's eye.

For the one-model problems, as we have remarked, the majority of conclusions were based on a single model. It is interesting to compare the models postulated in Polk and Newell's (1995) VR program with those constructed by our participants. As an example, consider again the one-model problem based on the premises:

Some B are A.

All B are C.

Under one interpretation of the premises, the VR program constructs the following model:

```
B′          C
B′   A      C
```

and then, as a result of re-encoding the first premise, it constructs the model:

```
B′          C
     A′
B′   A′     C
```

from which it generates the valid conclusion, Some A are C. The most frequent response (9 participants) in our experiment was to construct just the first of these models (ignoring the number of tokens). Of the five participants who constructed multiple models, two constructed the sequence above, although the first token constructed by all these participants was of the form B, A, C.

The sequences of models that the participants constructed showed that they again used the three operations embodied in the computer program, albeit in a striking variety of strategies. In addition, however, they also removed tokens from models, but almost always to revert to a model they had constructed earlier in the sequence. The sequences revealed three main sorts of error. The first sort of error was that participants constructed only one

of the possible models of the premises and based their conclusion on this model. For example, given the premises:

None of the A are B

Some of the C are B

a participant constructed the following model:

```
a
a
      b     c
      b
```

and drew the conclusion, None of the A are C. The failure to search for alternative models may result from an inadequate strategy or a working memory with a limited processing capacity.

The second sort of error was the failure to formulate the proper conclusion based on the models of the premises. For example, some participants constructed the model above, but then asserted that nothing followed from the premises. It is possible that these participants constructed an alternative model in their minds' eye:

```
a           c
a           c
      b     c
      b
```

but failed to grasp that a conclusion holds in both models, Some of the C are not A. We have noticed before that some poor reasoners tend to respond that "Nothing follows" whenever they construct more than one model of the premises (Johnson-Laird, 1983, p. 120–121). One explanation of this response is that these reasoners assume without further thought that no valid conclusion exists because they were able to construct more than one model. Another explanation is that these reasoners are unable to put into words, particularly into a singly quantified assertion, what is common to the models that they have constructed from the premises. This idea led Rips (1994) to use the experimental procedure in which the participants evaluated *given* conclusions. It is also supported by Greene (1992), who showed that his participants had difficulty in describing given models with certain multiply quantified assertions.

The third sort of error was that reasoners constructed a set of possible models of the premises, but then based their conclusion on only one of them. Given the following premises, for example:

Some of the A are B

None of the C are B

a participant constructed an external model of this sort:

```
a     b
a
      b
            c
```

She next added a token of C to create the model:

a b
a c
b
c

But, she drew the conclusion:

Some of the A are C

which holds only for the second model. The most plausible explanation for this error is that reasoners forget about their first model and what conclusion, if any, it supports.

The participants made errors of all three sorts, and none of them made errors of just one sort. The most frequent sort of error (made by 11 participants) was the failure to construct an alternative model of the premises. Four participants failed to draw the correct conclusion for any of the multiple-model syllogisms. Three of them tended to construct only one model of the premises; and one of them responded "No valid conclusion" in every case, even though she had constructed correct alternative models of the premises. In the previous experiment, we noted that the participants were often confused about what counted as a refutation of a premise in the O mood (e.g., "Some of the A are not C"). Judging from the remarks made by two of the participants, an analogous confusion occurred in the present experiment. These two participants said that "Nothing follows" if they could infer both an I conclusion and an O conclusion from the premises.

Discussion

The experiment showed that individuals not trained in logic do tend to construct multiple models in drawing their own conclusions from syllogistic premises. The evidence rests on examining the external models that they constructed in carrying out the task. We also examined the inferences that they made without the benefit of external models. They drew a less varied set of conclusions in this case. The difference arises, we believe, because the reasoners constructed external models without being able to encode negative information explicitly. They then drew their conclusion based on the model without considering the presence or absence of negative tokens. Hence, they were more likely to draw a negative conclusion from affirmative premises, or an affirmative conclusion from negative premises, than when they reasoned without the benefit of external models. For example, given the premises:

Some of the A are B

All the C are B

one participant constructed the model:

a b
a
b c

but then drew the conclusion:

None of the A are C

Yet, the participants' performance was of comparable accuracy whether or not they used external models, and in both of these conditions they drew reliably more correct conclusions to one-model syllogisms than to multiple-model syllogisms.

The evidence showed that naive reasoners construct a sequence of multiple models in drawing conclusions, especially from problems that support multiple models. This result was predicted by the model theory. Yet, the construction of multiple models, as a reviewer pointed out, is not necessarily equivalent to a search for counterexamples. A reasoner who constructs more than one model may be just augmenting an initial model, even when, as often happened, the reasoner constructs a model, modifies it, reverts to the original, modifies it again, and so on. Likewise, participants must in general have been envisaging an alternative model when they constructed a single model from which they drew a correct response that was inconsistent with that model. But, again, we cannot be certain that refutation was the underlying motivation. Hence, a sequence of models is suggestive evidence, but no more, for the claim that reasoners are searching for counterexamples. The explicit construction of multiple models on only 39% of trials might also seem to undermine the importance of falsification. On a further 14% of trials, however, the participants reached the correct conclusion, even though it was inconsistent with the one explicit model that they had constructed; a phenomenon that suggests that they had envisaged an alternative model. Moreover, when the participants failed to construct multiple models for those premises that supported them, they often drew erroneous conclusions.

There were two other common causes of error. Some errors occurred because reasoners constructed the right set of models but assumed that they had nothing in common; and some errors occurred because reasoners constructed the right set of models but described only one of them. To reach the right response to multiple-model problems for the right reason, it is necessary to consider not just their initial models, but to search for alternatives, to grasp what is common to all of them, and to describe it correctly.

V. GENERAL DISCUSSION

What have we learned from our studies? We will answer this question in two parts. The first part shows that no current theory of syllogistic reasoning can explain the results. The second part shows how a theory based on mental models can be developed to account for the phenomena.

In the Introduction, we described a variety of theories of syllogistic reasoning. It is impossible to prove that the processes postulated by a theory play no role in reasoning, and indeed many of these processes may occur. What we can show, however, is that a theory by itself cannot explain our results. The strategies of the participants in all four of our experiments demonstrate that logically untrained individuals are able to reason from syllogistic premises. They are not merely generating conclusions in accordance with the "atmosphere" of the premises (*pace* Wetherick & Gilhooly, 1990) or selecting a conclusion that matches the form of the least informative premise (*pace* Chater & Oaksford, 1999). Granted that individuals not trained in logic do reason, the principal controversy is whether they rely on formal rules of inference, some form of mental model, or both. Stenning and Yule (1997) argue that both sorts of theories can be subsumed within higher-order principles, at least in the case of syllogisms. They show that their algorithm

for Euler circles, which is isomorphic to one version of the model theory, is equivalent to a set of formal rules of inference. This framework is useful in accounting for certain aspects of syllogistic inference and of a task in which reasoners have to decide what individuals, if any, must exist given a pair of syllogistic premises. But, the framework is largely normative, accounting for how reasoners can reach correct conclusions. It makes no predictions about the sequences of models that reasoners construct or about the operations they use to generate such sequences. Indeed, sequences of alternative models play no part in the framework.

The view that some individuals use Euler circles, whereas others use formal rules, has been vigorously defended by Ford (1995). She carried out a study in which 20 members of the Stanford University community first "thought aloud" as they attempted to draw conclusions from the 27 pairs of syllogistic premises that yield valid conclusions, and then went through the same problems again explaining to the experimenter how they had reached their conclusions. Ford argued that the participants' protocols, diagrams, and explanations, enabled her to divide them two main groups, one using Euler circles and the other using verbal rules. She wrote "In contrast to the mental models theory given by Johnson-Laird and his colleagues, neither group makes use of representations containing finite elements standing for members of sets." (p. 19). However, her results did confirm that, in our terms, one-model problems were easier than multiple-model problems. Her verbal rules are a special case of a theory based on formal rules of inference, and we turn now to an assessment of such theories.

Formal Rules of Inference

Braine and Rumain (1983) and Rips (1994) proposed formal rules of inference for reasoning with quantified assertions. For example, Rips (1994) proposes a number of such rules, including the following one:

All X are Y.
All Y are Z.
∴ All X are Z.

The drawback of these rule systems is that models play no part in them, and so they are unable to explain the performance of the participants in our experiments. However, Rips's (1994) system has the power of a Universal Turing machine, and so it can be used as a programming language in which to implement any theory, including the mental model theory. His theory is thus almost irrefutable; that is, no empirical results could ever show it to be false unless they demonstrated that mental processes are not computable (Johnson-Laird, 1997).

Could it be that some reasoners do rely on formal rules? This claim, as we have seen, is defended by Ford (1995). She argued that eight of the participants in her study relied on a verbal substitution strategy. She classified participants as using this strategy if they spoke of replacing one term in a syllogism with another, or crossed out one term and replaced it with another. But, she also classified participants as using the strategy if they rewrote a syllogism as an equation or drew arrows between its terms (see Ford, 1995,

footnote 2, p. 18). This evidence may be consistent with the strategy, but it is hardly decisive. Consider how the strategy is supposed to work: "the subjects. . . take one premise as having a term that needs to be substituted with another term and the other premise as providing a value for that substitution" (Ford, 1995, p. 21). She proposed four principles that are supposed to govern substitutions. The first of them reads as follows (p. 21):

> If a rule [i.e., a premise] exists affirming of every member of the class C the property P, then whenever a specific object, O, that is a member of C is encountered it can be inferred that O has the property P.

The phrase "a specific object, O" refers to either "some of the O" or to "all of the O". In other words, reasoners are equipped with the following pair or rules of inference:

i)	All the C are P.	All the C are P.
	Some O are C.	All the O are C.
	∴ Some O are P.	∴ All the O are P.

We can similarly translate Ford's (1995) other three principles into pairs of rules, where the same quantifier must occur in both the second premise and the conclusion:

ii) All the C are P.
All/Some O are not P.
∴ All/Some O are not C.

iii) None of the C is P.
All/Some O are C.
∴ All/Some O are not P.

iv) None of the C is P.
All/Some O are P.
∴ All/Some O are not C.

Apart from notational differences, Braine and Rumain (1983) proposed identical rules. Not all valid syllogisms can be derived by using Ford's (1995) four principles, and so she goes to some pains to show how more sophisticated substitutions can be made where necessary. The strongest point of her paper is that the more sophisticated substitutions yield reliably poorer performance by those participants whom she classified as using the procedure, but not by those whom she classified as using Euler circles.

But, is the substitution procedure a purely verbal one dependent on formal rules of inference? And does the model theory, as Ford (1995) implies, group "all people together as though they basically reason in the same fashion" (p. 3)? Johnson-Laird and Bara (1984) wrote, "There are undoubtedly differences from one individual to another in the way in which they make syllogistic inferences. Our alternative implementations of the theory suggest a way in which some of these differences might be explained." (p. 50). Ironically, one of these alternatives was a substitution procedure based on models rather than verbal premises. This procedure was described in the following terms by Johnson-Laird (1983, p. 106):

> With premises in the A - B, B - C figure [arrangement], the two instances of the middle term, B, occur one after the other, and it is easy to construct a mental model of the first

> premise and then immediately integrate the information from the second premise. For example, with premises of the form:
>
> Some of the A are B
>
> All of the B are C
>
> a reasoner can form a model of the first premise:
>
> a = b
> a = b
> (a) (b)
>
> and then immediately integrate the content of the second premise by substituting Cs for Bs:
>
> a = c
> a = c
> (a) (c)
>
> This procedure for substituting one type of token for another. . . is an essential part of the explanation of the figural effects. . .

In fact, no evidence shows that the substitution strategy depends on verbal rules as opposed to mental models. Participants in our experiments who represented the premises:

All the A are B.

All the B are C.

in either of the following ways:

$A \rightarrow B \rightarrow C$ $A = B = C$

would be classified by Ford (1995) as using the verbal substitution strategy. But, there is no reason to believe that arrows or equalities are the outward signs of inward verbal substitutions. The participants could just as well be making substitutions in models, or indeed not making substitutions at all (see the alternative algorithm described by Johnson-Laird and Bara, 1984, p. 30). Even the cases where a participant replaced one word by another could reflect a substitution in a model. The difference between Ford's (1995) two groups of participants may reflect whether or not individuals use a substitution strategy. But, the use of diagrams and words in our experiments, which is comparable to their use in Ford's (1995) experiment, is neutral about whether substitutions concern tokens in models or constituents in sentences.

There are two further weaknesses in current formal rule theories. First, because models play no part in them, reasoners can respond, "No valid conclusion", only when they have failed to find a derivation of a conclusion. Unlike the participants in Experiment 3, they should not be able to refute a conclusion by constructing a counterexample to it. Second, current formal rule theories offer no account of systematic errors. They allow, of course, that reasoners can err, by misapplying a rule, for example, or by failing to find a derivation. The theories predict that such errors are more likely to occur with inferences that call for a greater number of inferential steps or that call for the use of more complicated rules (see e.g., Rips, 1994, Ch. 11). But, what they cannot predict is the nature of the particular errors that will occur. They should be haphazard rather than systematic. In fact, the majority of erroneous conclusions in our experiments were predicted by the model theory because they correspond to conclusions based on a single model of

multiple-model problems. It is odd that early theories of syllogistic reasoning emphasize the causes of error (e.g., Woodworth & Sells, 1935), whereas recent formal rule theories emphasize correct conclusions. A good theory ought to explain both.

Euler Circles

An alternative hypothesis about mental models for syllogisms is that they take the form of Euler circles. Their traditional use calls for the construction of all the different diagrams for each premise and all the different combinations for the pair of premises—a demand that leads to a combinatorial explosion (see e.g., Erickson, 1974). Stenning et al. (see e.g., Stenning & Oberlander, 1995), as we have seen, have devised a novel way to use Euler circles that obviates this explosion. Ford (1995) postulates a similar procedure: reasoners assume that areas enclosed by circles can be empty, and they use the verbal premises as reminders of which areas cannot be empty. This procedure, as Ford (1995) shows, is equivalent to the use of optional elements in models. Hence, the main burden of her results, and ours, is that reasoners who use Euler circles avoid the traditional method, but instead use methods closely resembling those that we have proposed for mental models.

There remain three questions about Euler circles. The first question is whether individuals who have never seen circles used to represent sets—either in logic or the so-called "new math"—spontaneously use Euler circles. Ford's (1995) participants were sophisticated, and some of them refer explicitly to "classes", "intersections," and the like, which suggests that they had seen circles used to represent sets. Our participants may have had similar encounters at school, although they claimed not to have been taught Euler circles. Perhaps they did re-invent the circles, but we are skeptical. As far as we know, no logician before Leibniz used circles to represent sets. The idea was a major innovation, and it was later popularized by Euler's letters to a Swedish princess. If naive individuals spontaneously use the method, why was it not invented earlier, and why did it have to be popularized?

The second question is whether people who draw Euler circles rely on visual images of them when they are denied paper and pencil. In other words, are Euler circles the natural way in which these individuals represent the extensions of quantified assertions? If circles were totally natural as mental representations, reasoners should have little need to draw them unless their externalization reduces the load on working memory. But, these drawings do not yield any great improvement in performance. Amongst our participants, those who used them did not perform reliably better than those who did not use them. And, unlike our format for mental models—finite numbers of tokens that represent individuals, Euler circles do not generalize beyond syllogisms to relational inferences. Individuals not trained in logic can reason from premises containing a mixture of assertions from syllogisms and relational assertions, but it seems implausible that they shift from one form of mental representation to another as they work their way through the premises.

The third question is what mental representations people have in mind when they draw Euler circles. One hypothesis is again that they have a visual image of an Euler circle. But, it is not the only possibility. An individual may have a more abstract representation in mind that controls the process of drawing. We make no strong claims for such an abstract representation. Ford (1995), however, appears to take for granted that because some of her participants drew Euler circles, it follows that these individuals were not using mental models of the sort that we have postulated. She writes, "Thus, the spatial subjects used a type of representation specifically dismissed by Johnson-Laird and his colleagues, where the class itself and not the finite members of the class is represented." (p. 41). Readers should note the equivocation in this claim. Ford (1995) is referring to the external representations drawn by her participants; Johnson-Laird et al. were referring to internal mental representations. Moreover, contrary to Ford (1995), some of the participants who were classified as verbal reasoners in her experiment do refer to individual entities, as the following extracts from four of her protocols show:

> ". . . if there are any historians like suppose there's two historians right that means there are two weavers who are also historians so we can say some of the weavers are historians. . . " (Eric)
>
> ". . . could have a weaver that is not a historian and is a TC member" (Catherine)
>
> ". . . all of the historians are weavers none of the historians well you actually can't conclude that because you have another some one else like a philosopher who could be a weaver who might be a tennis club member. . . " (Hilary)
>
> ". . . if you're a playwright you're always a bookworm that means you have a chance to be a stamp collector. . . " (Amy)

In short, some individuals sometimes draw Euler circles when they make syllogistic inferences. We are inclined to Rips's (1994) view that they rely on a vestigial memory for a procedure that they encountered in school. Euler circles, however, are a legitimate hypothesis about the nature of mental models. We do not know whether those who draw Euler circles use visual images of them either to control their drawings or to reason when they have no access to paper and pencil. But, we do know that they are not powerful enough for reasoning with relational premises, and that current psychological theories based on them cannot account for the results of Experiments 3 and 4.

Polk and Newell's Verbal Reasoning Theory

Polk and Newell (1995) propose that syllogistic reasoning depends on encoding and re-encoding premises as mental models rather than on a search for counterexamples. They support their claim by showing that "falsification" yields little improvement in the fit of VR to the data. We suspect that there is little improvement because VR does some of the work of refutation in other ways. What is right about their theory, however, is its emphasis on the variety of different interpretations of the premises. What may be wrong are the successive re-encodings of the premises. Figure 5 in Polk and Newell (1995, p. 539)

shows the construction of two distinct models for what, for our program, is a one-model problem. If this performance is typical of VR, then it contrasts with the performance of our participants. They tended to construct multiple models, not for one-model problems, but for multiple-model problems. Experiment 3 shows that reasoners are able to search for counterexamples if they are asked to do so. Experiment 4 is consistent, but hardly decisive. There is no doubt that the participants generated sequences of models, but whether they were searching for counterexamples is unclear. However, one other phenomenon in syllogistic reasoning does support a search for counterexamples. Byrne and Johnson-Laird (1990) tested their participants' ability to recognize conclusions that they had drawn earlier to syllogistic premises. As a search for counterexamples predicts, they often falsely recognized conclusions supported by an initial model of the premises when, in fact, they had responded correctly that nothing followed from the premises. This phenomenon suggests that they had fleetingly considered the erroneous conclusion only to reject it as a result of a counterexample.

With hindsight, we see that syllogisms are not an ideal test case for demonstrating a search for counterexamples. Modal reasoning is better because the model theory predicts an obvious interaction that hinges on reasoners searching for counterexamples: it should be easier to determine that a situation is possible (one model suffices) than necessary (all models must be checked), whereas it should be easier to determine that a situation is not necessary (one model serving as a counterexample suffices) than not possible (all models must be checked for counterexamples). The interaction has been corroborated in reasoning both from sentential connectives (Bell & Johnson-Laird, 1998) and in a recent unpublished study of quantified assertions carried out by Jonathan Evans et al. (see also Galotti, Baron, and Sabini, 1986). Hence, in our view, Polk and Newell (1995) may be right about syllogisms, but, in those tasks where counterexamples are of obvious use, reasoners appear to search for them. Moreover, as Barwise (1993) emphasized the only way one can *know* that a conclusion is invalid is by constructing a model of the premises that is a counterexample to it.

The Theory of Mental Models

Our experiments have shown that individuals not trained in logic try to reason when they draw conclusions to syllogistic premises, that they are able to refute conclusions by constructing counterexamples, and that they may do so spontaneously when they construct external models in order to draw their own syllogistic conclusions. These results are compatible with the idea that people reason by envisaging the situations that premises describe; that is, by constructing models of these situations. However, the program implementing the model theory is wrong. It departs from human performance in several important ways. First, reasoners do not adopt a fixed interpretation for each sort of syllogistic premise, but instead their interpretations vary in ways that are not entirely predictable (cf. Polk & Newell, 1995). Second, reasoners use a variety of strategies in reasoning; a phenomenon that we have also observed in a sentential reasoning study

carried out in collaboration with Fabien Savary (unpublished data). In the present experiments, the participants differed in which premise they interpreted first, in how they interpreted the premises, and in how they went about searching for alternative models. Their strategies were much more variable than we had envisaged in implementing the program. They used different sequences of operations to reach the same result (or different results). They differed one from another and from one problem to another. They generally realized when a conclusion could not be refuted, and they could construct counterexamples to those conclusions that had them. They constructed a model of the premises, which typically satisfied the conclusion—the putative conclusions were chosen on the basis of frequent errors—and then they tried to refute the conclusion by adding new individuals, by breaking individuals into two, or by joining two individuals to make one. This pattern of results is in accord with the mental model theory. Although we have hitherto disavowed the reality of the search operations used in our syllogistic algorithms, they do correspond closely to those that the participants used, at least in constructing external counterexamples. Where the protocols diverge radically from our current computer program is in their variety.

Certain aspects of the participants' performance are predictable in a probabilistic way. For example, they usually began with the first premise; they usually made a co-extensive interpretation of a premise of the form, All the A are B; and they usually used the operation of adding new individuals to models to search for alternative models. But, it was impossible to predict precisely what an individual would do on a particular trial. Hence, there is no option but to build an element of non-determinism into a theory of reasoning. In our view, individuals seldom carry out a fixed deterministic strategy in any sort of thinking, whether it is deductive reasoning, inductive reasoning, creating new ideas, or daydreaming (Johnson-Laird, 1991). One way to express such a theory is as a grammar with alternative rules that allow for alternative ways to represent premises, formulate conclusions, and search for counterexamples. In this way, the theory could be used to "parse" each sequence of models constructed in reaching conclusions. It is now clear why the programs implementing the model theory of syllogisms have kept changing over the years. The programs have tried to account in a simple, deterministic way for performance that cannot be shoehorned into such a narrow confine.

Granted the diversity of strategies, what are the main components of syllogistic competence, and what are the robust phenomena of performance? Competence includes a number of core abilities: naive reasoners are able to understand quantified assertions; to envisage the situations in which they are true; and to construct external models of them, either as Euler circles or as those models that have finite numbers of individual tokens. They can use such representations to describe necessary, possible, and impossible individuals, and to formulate necessary, possible, or impossible quantified conclusions. Given a putative conclusion, they can refute it if called upon to do so. This list of competencies transcends all current theories of syllogistic reasoning. We believe that it is best accounted for in terms of a semantic theory that explains how individuals can map assertions into models, and vice versa. A key principle at the heart of the model theory is a conclusion is necessary—it must be true—if it holds in all the models that satisfy the premises; a

conclusion is probable—it is likely to be true—if it holds in most of the models that satisfy the premises; and a conclusion is possible—it may be true—if it holds in at least one model that satisfies the premises. Likewise, a particular sort of individual is necessary if there is a token of such an individual in all the models that satisfy the premises, a particular sort of individual is probable if there is a token of such an individual in most of the models that satisfy the premises, and a particular sort of individual is possible if there is a token of such an individual in at least one model that satisfies the premises.

Performance reveals some robust phenomena. First, individuals differ enormously in their syllogistic ability. In earlier studies, we have observed at least one adult not trained in logic who drew 85% correct conclusions, and at least one who drew only 15% conclusions (see Johnson-Laird, 1983, p. 118–9). The present study confirmed such differences. Why do they exist? As the model theory predicts, the capacity of working memory and the ability to perceive what is common to two situations account for part of the variance (Bara et al., 1995). Otherwise, we are far from answering this important question.

Second, syllogisms themselves vary enormously in difficulty. The mental model theory captures a general truth about this difference. We can contrast this truth with another characteristic of syllogisms. All syllogisms are referentially indeterminate. Thus, a syllogism of the form:

Some of the A are B.

All the B are C.

can refer to a variety of different states of affairs in the universe of discourse. It may refer to a universe in which there is only one sort of individual:

a b c

or to a universe in which there are two sorts of individual:

a b c

a −b c

or to a universe in where there are still more kinds of individual. In general, the existence of the following different kinds of individuals is consistent with the premises:

a b c

a −b c

a −b −c

−a b c

−a −b c

−a −b −c

Only the first sort of individual, however, necessarily exists. The others may or may not exist. Hence, the premises are consistent with $2^5 = 32$ distinct states of affairs. Here is the crucial point: the premises call for just one mental model (see e.g., Table 4), and the problem is indeed very easy.

In contrast, consider the following premises:

Some of the A are B

None of the B are C

They, too, are consistent with the existence of the following individuals:

a	b	−c
a	−b	c
a	−b	−c
−a	b	−c
−a	−b	c
−a	−b	−c

Only the first sort of individual and at least one individual that is a C must exist, granted the existence of As, Bs, and Cs. Hence, the premises are consistent with 24 distinct states of affairs. In this case, however, the premises call for multiple models. This syllogism is difficult, and many reasoners draw the invalid conclusion corresponding to the first model constructed by the computer program.

The general principle governing the difficulty of syllogisms is straightforward: those syllogisms for which the valid conclusion depends on one model are easier than those for which the valid conclusion depends on more than one model. This difference survives the great variety of strategies and interpretations that our experiments have demonstrated. This diversity can be captured only by a non-deterministic process that allows alternative search routes to be taken at almost any point in the process. Over the years, our tinkering with the implementation of the model theory was a vain attempt to capture the highly flexible human system of reasoning within one deterministic framework. The general principle governing the diversity of strategies is that reasoners are searching for alternative models of the premises. The search may be an attempt to examine all the possibilities, or it may be guided by the goal of refuting putative conclusions. In either case, the plausibility of different psychological theories depends on how closely they reflect two phenomena, the generation of alternative models and the difference between one-model and multiple-model syllogisms. They can be captured by theories that postulate mental models, even models in the form of Euler circles, but not by other current theories.

Acknowledgments We thank many colleagues for their advice, especially Bruno Bara, Ruth Byrne, Fabien Savary, and Yingrui Yang. We are also grateful to Clayton Lewis and two anonymous referees for their helpful comments on an earlier version of the paper. The research was supported in part by ARPA (CAETI) contracts N66001–94-C-6045 and N66001–95-C-8605.

REFERENCES

Bara, B., Bucciarelli, M., & Johnson-Laird, P. N. (1995). The development of syllogistic reasoning. *American Journal of Psychology, 108,* 157–193.

Barwise, J. (1993). Everyday reasoning and logical inference. (Commentary on Johnson-Laird and Byrne, 1991) *Behavioral and Brain Sciences, 16,* 337–338.

Bell, V., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science, 22,* 25–51.

Braine, M. D. S., & Rumain, B. (1983). Logical reasoning. In J. H. Flavell & E. M. Markman (Eds.), *Carmichael's handbook of child psychology, Vol. III: Cognitive development* (4th ed., pp. 263–339). New York: Wiley.

Byrne, R. M. J., & Johnson-Laird, P. N. (1990). Remembering conclusions we have inferred: What biases reveal. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases: Their contribution for understanding human cognitive processes* (pp. 109–120). Amsterdam: North-Holland.

Cardaci, M., Gangemi, A., Pendolino, G., & Di Nuovo, S. (1996). Mental models vs. integrated models: Explanations of syllogistic reasoning. *Perceptual and Motor Skills, 82,* 1377–1378.

Chapman, L. J., & Chapman, A. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology, 58,* 220–226.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38,* 191–258.

Erickson, J. R. (1974). A set analysis theory of behaviour in formal syllogistic reasoning tasks. In R. Solso (Ed.) *Loyola symposium on cognition* (Vol. 2, pp. 305–329). Hillsdale, NJ: Lawrence Erlbaum Associates.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, East Sussex, England: Psychology Press.

Fisher, D. L. (1981). A three-factor model of syllogistic reasoning: The study of isolable stages. *Memory and Cognition, 9,* 496–514.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54,* 1–71.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology General, 115,* 16–25.

Greene, S. (1992). Multiple explanations for multiply quantified sentences: Are multiple models necessary? *Psychological Review, 99,* 184–187.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41–82). New York: Seminar Press.

Guyote, M. J., & Sternberg, R. J. (1981). A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology, 13,* 461–525.

Hardman, D. K. (1996). Mental models: The revised theory brings new problems. *Behavioral and Brain Sciences, 19,* 542–543.

Henle, M., & Michael, M. (1956). The influence of attitudes on syllogistic reasoning. *Journal of Social Psychology, 44,* 115–127.

Johnson-Laird, P. N. (1975). Models of deduction. In R. J. Falmagne (Ed.), *Reasoning representation and process in children and adults* (pp. 7–54). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Johnson-Laird, P. N. (1991). *Human and machine thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N. (1997). Rules and illusions [A critical study of Rips's *The Psychology of Proof*]. *Minds and Machines, 7,* 387–407.

Johnson-Laird, P. N., & Bara, B. (1984). Syllogistic inference. *Cognition, 16,* 1–61.

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

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64–99.

Martín-Cordero, J., & González-Labra, M. J. (1994). Amnesic mental models do not completely spill the beans of deductive reasoning. *Behavioral and Brain Sciences, 17,* 773–774.

Newell, A. (1980). Reasoning, problem solving and decision processes: The problem space as a fundamental category. In R. Nickerson (Ed.), *Attention and performance* (Vol. 8, pp. 693–718). Hillsdale, NJ: Lawrence Erlbaum Associates.

Oakhill, J. V., & Johnson-Laird, P. N. (1985). The effects of belief on the spontaneous production of syllogistic conclusions. *Quarterly Journal of Experimental Psychology, 37A,* 553–569.

Oakhill, J. V., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31,* 117–140.

Polk, T. A. (1993). Mental models, more or less. *Behavioral and Brain Sciences, 16,* 362–363.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102,* 533–566.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior, 14,* 180–195.

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

Russell, B. A. W. (1946). *History of western philosophy*. London: Allen & Unwin.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science, 19,* 97–140.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology, 34,* 109–159.

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse. *Archiv für die gesamte Psychologie, 11,* 1–27.

Wetherick, N. E., & Gilhooly, K. J. (1990). Syllogistic reasoning: Effects of premise order. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking* (Vol. 1, pp. 99–108). New York: John Wiley.

Woodworth, R. S. & Sells, S. B. (1935). An atmosphere effect in formal reasoning. *Journal of Experimental Psychology, 18,* 451–460.

Yang, Y., & Johnson-Laird, P. N. (1997). Illusions in quantified reasoning How to make the impossible seem possible, and vice versa. *Memory & Cognition,* in press.